OPEN ACCESS Freely available online

ISSN: 2090-4908
International Journal of Swarm Intelligence and Evolutionary Computation

Research Article

Hybrid Model using LOF and iForest Algorithms for Detection of Insider Threats

Anvita Ajay Mahajan*

Department of Computer Science, COEP Technological University, Pune, India

ABSTRACT

Insider threats, is one of the most challenging threats in cyberspace, usually responsible for causing significant loss to organizations. The topic of insider threats has long been studied and many detection techniques were proposed to deal with insider threats. This paper focuses on using different anomaly detection algorithms locality outlier factor algorithm and Isolation forest Algorithm and does a comparative analysis between their performances. A hybrid model incorporating advantages of both LOF Algorithm and IF Algorithm is proposed in this paper which gives better performance than the individual models for detecting insider threats. The hybrid model was able to achieve whooping 99.99% accuracy while detecting insider threats.

Keywords: Insider threats; Locality outlier factor; Isolation forest; Hybrid models; CERT dataset

INTRODUCTION

Insider threats account for the lion's share of the risk faced by businesses and institutions. Insider threats are more damaging to an organization than outsider/intruder attacks, since there are significant costs associated in mitigating insider attacks. The workshop also provided a working definition of insider threat "Any authorized user who performs unauthorized actions that result in loss of control of computational assets". A 2022 report by the Indian Computer Emergency Response Team (CERT-In) found that insider threats accounted for 22% of all cyber-attacks in India [1]. Industry surveys show that 79% of security threats are insider threats, *i.e.*, malicious acts carried out by the organisation's careless or disgruntled employees who abuse their authorised access to networks, systems, and data [2]. A good starting definition was presented at the Rand 2000 workshop on insider threats to information systems "Any authorized user who performs unauthorized actions". This definition is more suited to the Information Technology (IT) domain in general and to the risk analysis and risk mitigation process specifically [3].

So basically insider threat is a malicious activity against an organization that comes from users with legitimate access to the organizations networks, applications or databases. These can be current employers, former employers, vendors, con-tractors, or partners. Insider threats can be broadly classified into three types:

- Malicious-Act done knowingly.
- Negligent-By someone who does not follow IT procedures.
- Compromised-Someone whose credentials are unknowingly compromised.

It is really challenging and difficult task to detect a threat and mark it as suspicious as it is done by the users who have access to the system. Also, another issue is lack of real time data available and the currently available dataset is highly imbalanced.

In this paper considering insider threat detection to be a specific use case of anomaly detection which is a general technique used to identify unusual or irregular patterns in the data, a hybrid is created using Isolation Forest algorithm and Locality Outlier Factors (LOF). This hybrid model leverages the strengths of both LOF and isolation forest. LOF is good at identifying local anomalies and is sensitive to data density, while isolation forest is effective at finding global anomalies and isolates them efficiently. By combining these models, the overall effectiveness of the insider threat detection system is tremendously improved.

Correspondence to: Anvita Ajay Mahajan, Department of Computer Science, COEP Technological University, Pune, India; E-mail: mahajananvita3@gmail.com

Received: 17-Nov-2023, Manuscript No. SIEC-23-23935; **Editor assigned:** 20-Nov-2023, PreQC No. SIEC-23-23935 (PQ); **Reviewed:** 06-Dec-2023, QC No. SIEC-23-23935; **Revised:** 01-Apr-2025, Manuscript No. SIEC-23-23935 (R); **Published:** 09-Apr-2025, DOI: 10.35248/2090-4908.25.14.413

Citation: Mahajan AA (2025) Hybrid Model using LOF and iForest Algorithms for Detection of Insider Threats. Int J Swarm Evol Comput. 14:413.

LITERATURE REVIEW

Research on insider threat detection has been ongoing for many years. Some of the early research focused on developing rule based systems that could identify known patterns of malicious activity. However, as insider threats have become more sophisticated, rule based systems have become less effective. In recent years, researchers have begun to focus on developing more sophisticated ITDSs that use machine learning and Artificial Intelligence (AI) techniques. These systems can be more effective at identifying insider threats because they can Learn and adapt to new patterns of malicious activity [4].

Here is a review of some of the key research papers on insider threat detection

- A hybrid model for insider threat detection using rule based detection and support vector machines: This paper proposes a hybrid ITDS that combines rule based detection with Support Vector Machines (SVMs). The system was able to achieve an accuracy of 99% and a false positive rate of 0.29% for known insider threats [5].
- A hybrid model for insider threat detection based on behavioral anomaly detection and random forest Algorithm: This paper proposes a hybrid ITDS that combines Behavioral Anomaly Detection (BAD) with a random forest algorithm. The system was able to achieve an accuracy of 97% and a false positive rate of 2.88% for unknown insider threats [6].
- A hybrid model for insider threat detection in cloud computing environment using rule based detection and deep learning algorithm: This paper proposes a hybrid ITDS that combines rule based detection with a deep learning algorithm. The system was able to achieve an accuracy of 98.5% and a false positive rate of 0.05% for detecting insider threats in a cloud computing environment [7].
- A hybrid model for insider threat detection in social network environment using behavioural anomaly detection and graph based algorithm: This paper proposes a hybrid ITDS that combines BAD with a graph-based algorithm. The system was able to achieve an accuracy of 97.2% and a false positive rate of 1.89% for detecting insider threats in a social network environment [8].
- A hybrid model for insider threat detection in distributed systems using user and entity behaviour analytics and game theory-based algorithm: This paper proposes a hybrid ITDS that combines User and Entity Behaviour Analytics (UEBA) with a game theory-based algorithm. The system was able to achieve an accuracy of 96.8% and a false positive rate of 0.93% for detecting insider threats in a distributed system environment.
- Insider threat detection based on user behaviour modelling and anomaly detection algorithms. Four anomaly detection algorithms used in this paper. (a) Gaussian density estimation, (b) Parzen window density estimation (reprinted from Alpaydin), (c) Principal Component Analysis (PCA), and (d) K-Means Clustering (KMC) and their combinations. It was observed that these work well on unbalanced dataset too [9].

These are just a few examples of the many research papers that have been published on insider threat detection. The research in this field is ongoing, and new and more sophisticated methods are being developed all the time.

METHODOLOGY

Dataset

CERT dataset which is one of the commonly used datasets for insider threats detection is used here. This dataset was developed by CERT division of the software engineering institute at carnegie mellon university. CERT dataset consists of log files out of which 5 used here are that record the computer based activities for all employees in an organization, including logon.csv that records the logon and logoff operations of all employees, email.csv that records all the emails and its details, ttp.csv that records all the web browsing, file.csv that records activities involving a removable media device, and device.csv that records the connect and disconnect timings usage of a thumb drive.

There are several versions of datasets according to when the datasets were created. Here, r4.2 version of CERT Dataset is used as it is a dense dataset containing many insiders and malicious activities [10].

Data analysis and feature extraction

The necessary features for training data are extracted. 28 different user actions were identified as features. They are shown in the Table 1 below.

Table 1: Features extracted from user activities.

File	Features
email.csv	• Email (checks if email) • Mass email from insider • Email from insider which is not mass • Mass email from outsider • Email from outsider which is not mass
logon.csv	• Connect • Disconnect • Connect during normal times (9 am-5 pm)

	• Connect after regular timings • connect at weekends
device.csv	• EXE file • JPG file • ZIP file • TEXT file • Documnet (DOC file) • PDF file
logon.csv	• Login • Logout • Weekday login between 9 am to 5 pm • Weekday login after decided timings • Weekend login
http.csv	• User visits social networking site • Cloud • Job sites • Data leak at mentioned website in readme. • Hacked website

Normalization

In order to ensure that different features in the dataset have the same scale, preventing some features from dominating others during machine learning model training, min-max normalization has been used throughout the research project. Minmax normalization [11]:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where

- $X_{\text{normalized}}$ is the rescaled value of the feature.
- X is the original value of the feature.
- $X_{\min}$ is the minimum value in the dataset for that feature.
- $X_{\max}$ is the maximum value in the dataset for that feature.

Models

Initially the results of Locality outlier factors and isolation forest were observed individually. After observing the results of individual models, a hybrid model was created for getting higher accuracy.

Isolation forest: Isolation forest is an algorithm for data anomaly detection initially developed by Fei Tony Liu in 2008 [12]. It internally uses binary trees. The isolation forest 'isolates' observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature [13]. The overall working of isolation forest is as summed up in Figure 1.

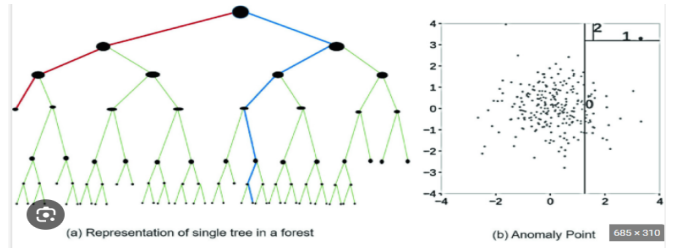

Figure 1: Isolation forest algorithm.

Internally it uses decision trees and a leaf for each and every data point is created. After this, anomaly score is calculated for every data point and a threshold value is determined. Generally, if the anomalyscore>threshold then the point is termed as anomalous.

Some of the advantages of using Isolation forest are:

- Isolation forest works very well when the sampling size is kept small, a property that is in contrast with the great majority of existing methods, where a large sampling size is usually desirable [11,12].
- Isolation forest performs well even if the training set does not contain any anomalous point, and the dataset used here is highly unbalanced hence this algorithm is highly compatible.

Locality outlier factor: In anomaly detection, the Local Outlier Factor (LOF) is an algorithm proposed by Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and J¨org Sander in 2000 for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbours [6]. The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbors. The overall working of isolation forest is as summed up in Figure 2.

It considers as outliers the samples that have a substantially lower density than their neighbors. Here the locality is given by

k-nearest neighbors whose distance is used to estimate density. This distance is termed as" Reachability distance".

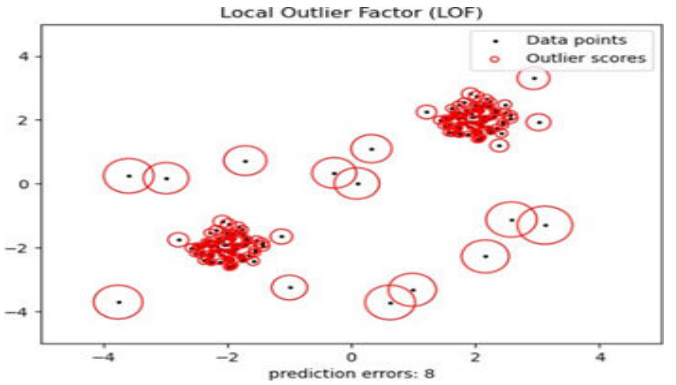

Figure 2: Locality outlier factor algorithm.

Following are some advantages of using LOF:

- Due to the local approach, LOF is able to identify outliers in a data set that would not be outliers in another area of the data set.
- While the geometric intuition of LOF is only applicable to low-dimensional vector spaces, the algorithm can be applied in any context dissimilarity function can be defined. Thus, this algorithm can be generalized easily [7].

Hybrid model: Despite the mentioned advantages, both the models suffer from various disadvantages like:

Isolation forest

- Deficiency in the identification of local anomalies, which affects the accuracy of the algorithm and results in onesided results [8].
- Performance with redundant data is not that good comparatively [1].

Locality outlier factor

- LOF requires a long execution time. [5]
- LOF is also sensitive to the minimum point's value which may result in false positives.

Thus, to overcome these disadvantages, a hybrid model is proposed. In this paper, the hybrid model is created by combining the results of the models obtained individually in a parallel manner and sequential manner. In parallel manner, the mean of anomaly scores of LOF and isolation forest is taken.

In serial manner, first the outliers are detected using isolation forest and later on the detected outliers are verified using LOF algorithm. Initially isolation forest algorithm is used because it is computationally efficient and LOF is later applied as it is able to identify the outliers locally.

RESULTS AND DISCUSSION

The dataset was split into 80% training data and 20% testing data. The testing data also consisted of the true positives data already provided in the dataset.

Isolation forest: Figure 3 shows overall distribution of anomaly scores obtained by isolation forest algorithm. The algorithm achieved training accuracy of 97.0953047844077% and testing accuracy of 89.953939816631O7%.

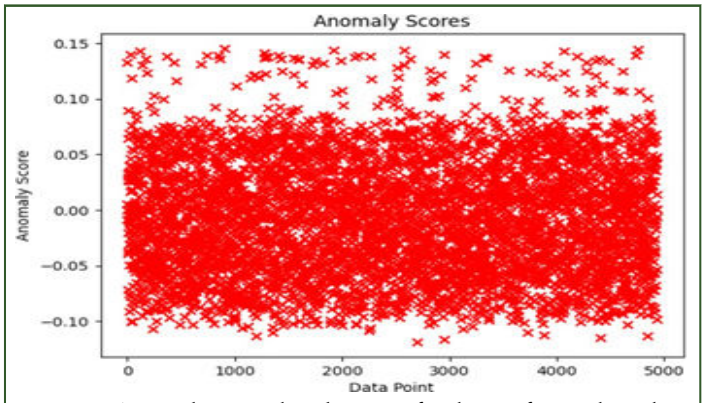

Figure 3: Anomaly score distribution of isolation forest algorithm.

Locality outlier factor: Figure 4 shows the overall distribution of outlier scores obtained while training by locality outlier factor algorithm. The algorithm achieved training accuracy of 99.91223295458356% and testing accuracy of 99.88867764466038%.

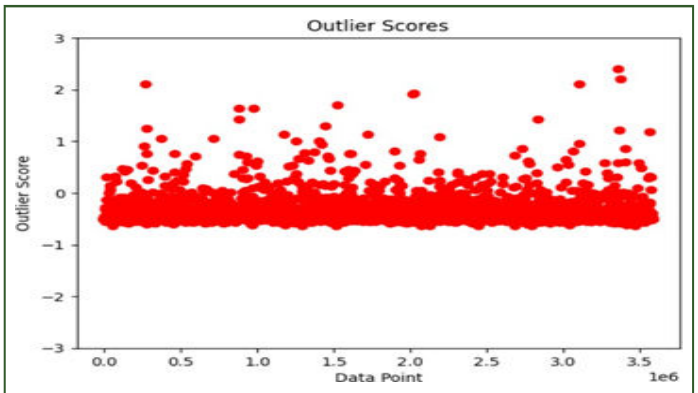

Figure 4: Outlier score distribution of locality outlier factor algorithm.

Hybrid model: Figure 5 shows the overall distribution of outlier scores obtained while training by locality outlier factor algorithm. This algorithm outperformed the individual models and achieved training accuracy of 99.9998602880525% and testing accuracy of 100%.

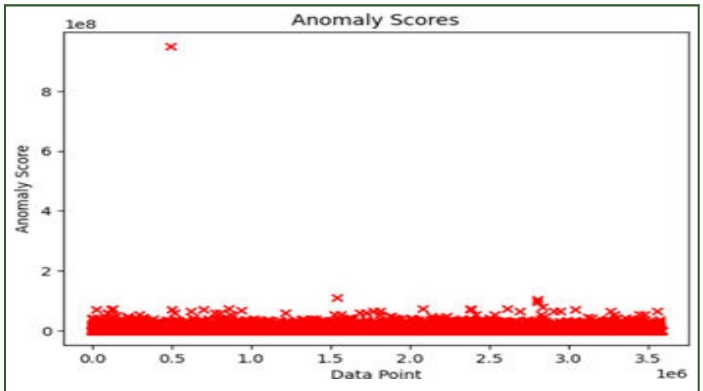

Figure 5: Hybrid model outlier scores.

The comparative analysis of two individual models and the hybrid model is shown in the Table 2 below:

Table 2: The comparative analysis of two individual models and the hybrid model.

Model	Precision	Recall	F1 score
IF training	99.99985%	97.09544%	98.52624%
IF testing	100%	90.01004%	94.74240%
LOF training	99.99986%	99.91237%	99.95609%
LOF testing	100.0%	99.88867%	99.94430%
Hybrid training	99.99986%	100.0%	99.99993%
Hybrid testing	100.0%	100.0%	100.0%

The Figures 6 and 7 show comparative analysis between the training and testing accuracies of the individual models and the hybrid model.

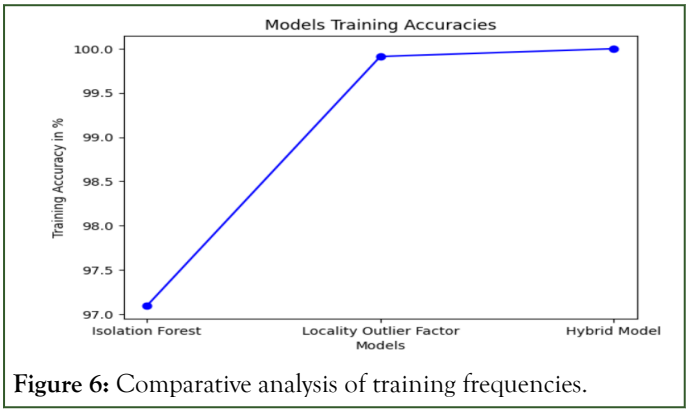

Figure 6: Comparative analysis of training frequencies.

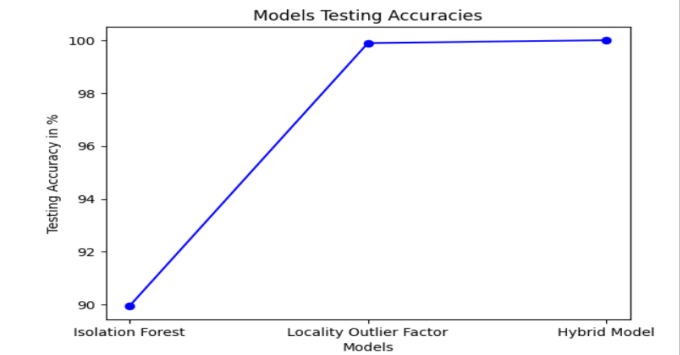

Figure 7: Comparative analysis of testing frequencies.

CONCLUSION

Thus, to conclude, the hybrid model tries to negate the disadvantages of individual models, and gives better results in identifying the insider threats. Although locality outlier factor algorithm and Isolation forest algorithm have been previously used for identifying insider threats, the hybrid model using them has not been researched upon. Similar to the hybrid model between locality outlier factor algorithm and isolation forest algorithm, in future we can create hybrid models between different anomaly detection algorithms available and do a comparative analysis between their results. For best results the models chosen for creating the hybrid model must be such that they nullify the disadvantages of the other chosen models.

REFERENCES

1. Mahajan A. Hybrid Model using LOF and iForest Algorithms for Detection of Insider Threats. Authorea Preprints. 2023.
2. Alghushairy O, Alsini R, Soule T, Ma X. A review of local outlier factor algorithms for outlier detection in big data streams. Big Data Cogn Comput. 2020;5(1):1.
3. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. InProceedings of the 2000 ACM SIGMOD international conference on Management of data 2000:93-104.
4. Campos GO, Zimek A, Sander J, Campello RJ, Micenkova B, Schubert E, et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Min Knowl Discov. 2016;30:891-927.
5. Gao R, Zhang T, Sun S, Liu Z. Research and improvement of isolation forest in detection of local anomaly points. Injournal of physics: conference series, New York, 2019 (Vol. 1237, No. 5, p. 052023). IOP Publishing.
6. Garg S, Kaur K, Kumar N, Kaddoum G, Zomaya AY, Ranjan R. A hybrid deep learning-based model for anomaly detection in cloud datacenter networks. IEEE Transactions on Network and Service Management. 2019;16(3):924-935.
7. Kim J, Park M, Kim H, Cho S, Kang P. Insider threat detection based on user behavior modeling and anomaly detection algorithms. Appl Sci. 2019;9(19):4018.
8. Liu FT, Ting KM, Zhou ZH. Isolation forest. In2008 eighth IEEE international conference on data mining, Pisa, Italy 2008 (pp. 413-422). IEEE.
9. Liu FT, Ting KM, Zhou ZH. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD). 2012;6(1):1-39.
10. Panja S, Patowary N, Saha S, Nag A. Anomaly Detection in IoT Using Extended Isolation Forest. InInternational Symposium on Artificial Intelligence. Cham: Springer Nature Switzerland. 2022:3-14.
11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Machine Learning Res. 2011;12:2825-2830.
12. Rahman MS, Halder S, Uddin MA, Acharjee UK. An efficient hybrid system for anomaly detection in social networks. Cybersecurity. 2021;4(1):10.

13. Ren X, Wang L. A hybrid intelligent system for insider threat detection using iterative attention. InProceedings of 2020 6th International Conference on Computing and Data Engineering. 2020:189-194).