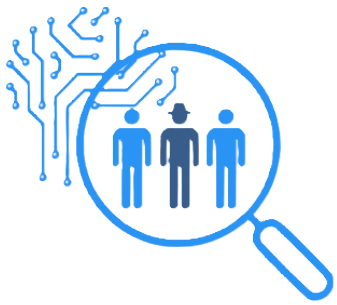


Project Report
InternalEYE: Insider Threat Detection Using ML

Written by:

Mona Alrayiqi
munamohamad1422@gmail.com

Raghad Alharbi
Raghad-alharbi1@outlook.com

Raghad Althagafi
Raghad7219@gmail.com

Reema Alharbi
Reema.226m@gmail.com

Bayadir Al-Luhaybi
bayader22k@gmail.com

Lara Aldughish
Laaldughish@gmail.com

Abstract:

Insider threats are a critical challenge to organizational security, often evading traditional detection methods. This study proposes a machine learning approach using Long Short-Term Memory (LSTM) networks to detect insider threats. By analyzing temporal patterns in user behavior, the LSTM model learns normal activity and identifies deviations as potential threats. The proposed method is evaluated on benchmark datasets, demonstrating improved detection accuracy and reduced false positives compared to traditional models. This approach offers a robust solution for enhancing the detection of insider threats in organizational environments.

1. Introduction

In today's world, almost all organizations face cyber attacks and thus focus on external protection. While measures like securely locking doors, controlling access with IDs, and monitoring with cameras enhance physical security, it's crucial to also address internal threats for comprehensive safety. True security requires a holistic approach to identify and mitigate risks, including those from internal sources, known as "**Insider Threats**."

An **Insider Threat** involves individuals within an organization who misuse their access privileges for malicious purposes, such as stealing information or seeking revenge. This can lead to significant data breaches, financial losses, and reputational damage [1]. These threats may arise from various factors, including malicious intent, negligence, or intentional actions by employees.

A 2024 Kaspersky report highlights that 71% of Saudi companies faced cyber attacks in the past two years, with 27% being insider threat attacks. This underscores the need to address these risks and protect organizational security [2].

This project addresses insider threats by developing InternalEYE, an AI solution designed to enhance the detection of anomalies in user behavior using advanced machine learning techniques, including Long Short-Term Memory (LSTM) and Dense neural network models [3]. InternalEYE proactively detects risks by analyzing user behavior to understand typical patterns and identify anomalies that may indicate malicious intent. It focuses on achieving high accuracy rates while minimizing false positives and negatives. The system utilizes sophisticated machine learning and behavioral analytics to monitor user activities, aiding security teams in detecting and responding to threats before they escalate. Additionally, InternalEYE features an intuitive dashboard, providing security teams with comprehensive visibility into potential threats and enabling swift responses to suspicious activities. This solution aims to create a multi-layered security framework that not only protects sensitive information but also promotes a culture of awareness and vigilance within the organization. Ultimately, the goal is to empower organizations to effectively combat insider threats and maintain a secure operating environment.

2. Problem Statement

The challenge of insider threats has become increasingly complex, as it is difficult to search for suspicious activity in all departments within organizations. While traditional security measures face significant challenges in detecting threats from trusted users, they often fail to identify malicious activities. This ineffective detection can lead to overlooked threats or false alarms, creating a false sense of security. Furthermore, the large amount of user behavior data complicates the identification of suspicious actions, making it increasingly difficult to protect organizations from insider threats.

Different from external hackers, insider threats start and arise from individuals already within the organization and part of it. When insider threats occur, the impact can be significant, sensitive information may be stolen, trust can be destroyed, and the organization's reputation suffers significant damage.

Employees with malicious intent are the most dangerous insiders, as they are unknown yet and their actions often go unnoticed and undetected. They possess detailed knowledge of the organization's infrastructure and processes, including familiarity with the information security tools in use.

Being already within the company's network, they do not need to penetrate external defenses from outside through methods like phishing or firewall attacks. Additionally, their relationships with colleagues make it easier for them to employ social engineering tactics.

Malicious insiders can severely disrupt business operations by sabotaging processes or leaking confidential information. This not only puts sensitive data at risk but also reveals weaknesses in the IT infrastructure, making it easier for outside attackers to exploit these vulnerabilities. Furthermore, the impact of insider threats extends beyond immediate financial losses. Organizations may suffer reputation damage, loss of customer trust, and legal repercussions.



Figure 1. The system architecture

3. System Architecture

The detection system is vital for organizations aiming to strengthen their defenses against insider threats. By quickly identifying and detecting potential internal risks, this system helps in preventing data breaches and safeguarding sensitive information. It allows organizations to maintain operational integrity, ensuring that business processes run smoothly without disruption. Additionally, by fostering a proactive security culture, the system encourages employees to be vigilant and aware of their role in safeguarding organizational assets. Ultimately, this system enhances overall security resilience, enabling organizations to navigate the complexities of modern threats effectively.

- **Insider Threat**

  Focuses on identifying potential malicious actions from individuals within the organization. It encompasses various behaviors that could indicate an insider threat, such as access to prohibited websites, unusual file downloads, or attempts to disclose sensitive information using thumb drives. Understanding these threats is essential for developing effective detection strategies.

- **Load Activity**

  The load activity component involves gathering and processing data from various sources within the organization. This includes monitoring user interactions, system logs, and network activities. By aggregating this data, the system creates a comprehensive view of user behavior, setting the stage for detailed analysis.

- **Analyze The Activity**

  Once the data is loaded, the system analyzes user activities to identify patterns and anomalies. This component employs advanced algorithms and machine learning techniques to assess normal behavior against potential threats. By analyzing activity, the system can detect deviations that may indicate malicious intent, enabling faster response times.

- **Detection**

  The detection component is responsible for confirming whether an identified anomaly constitutes a genuine threat. This involves applying predefined rules and machine learning models to evaluate the severity and credibility of the detected activity and whether the activities are classified as normal or abnormal. Accurate detection is critical to reducing false positives and ensuring that security resources are effectively utilized.

- **Alert**

  The alert component serves as the notification system for security teams. When a potential insider threat is detected, alerts are generated to inform the relevant personnel about the situation. These alerts can be customized based on the severity of the threat, allowing teams to prioritize their responses. Timely alerts are vital for quick action, minimizing the potential impact of insider threats.

-

4. Dataset

For this project, InternalEYE utilized the CERT Insider Threat Dataset, specifically version 4.2. This dataset is widely recognized in cybersecurity research for its realistic simulation of various insider threat scenarios within an organization [4]. It represents 12 months of activity logs for employees in a simulated environment, capturing a wide range of user activities including logons, logoffs, file accesses, emails, HTTP connections, and device usage. The dataset consists of five primary files, each focusing on a different aspect of user activities:

- Logon.csv: Records logon and logoff activities, including details such as user ID, timestamp, and the PC used.
- Device.csv: Captures device-related activities, such as the connection or disconnection of removable devices like USB drives.
- File.csv: Logs file access activities, including file creation, deletion, and modification.
- Email.csv: Contains records of email communications, detailing sender and recipient IDs, timestamps, and attachment information.
- HTTP.csv: Logs web browsing activities, including visited URLs, timestamps, and user IDs.

Each of these files includes a timestamp column, allowing us to track the sequence of user activities over time. Additionally, the dataset includes metadata about users, such as their roles within the organization, providing context for the activities captured in the logs.

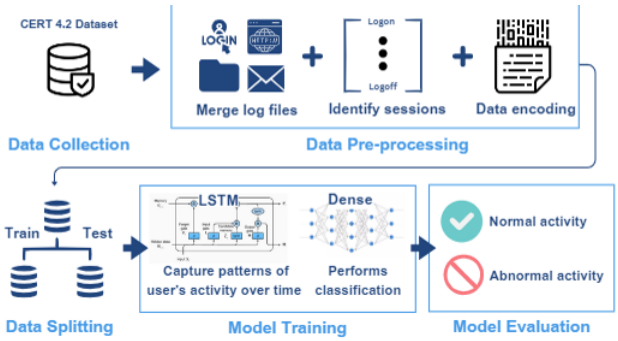


Figure 2. Methodology

5. Methodology

The methodology phase in Fig. 2 shows the general overview of the solution approach stages. highlighting the usage of the suitable dataset for this research. The figure describes the data pre-processing steps of the data for feeding it to the model. Then it focuses on model training using LSTM networks, which capture temporal patterns for detecting insider threats. It displays the evaluation process to measure model effectiveness, ensuring a rigorous investigation into insider threat detection.

5.1. Data Preprocessing

The data preprocessing phase is crucial for transforming raw log data into a format suitable for training the LSTM model. This phase consists of several key stages: Filtration, Merging, Session Extraction, and Encoding. Fig. 2 in the report illustrates the overall methodology used in this project, including the preprocessing steps.

1) **Loading and Filtration**: Given the large size of the CERT 4.2 dataset, efficient memory management is

essential. To address this, only relevant data from the log files are loaded. This initial filtration step is integrated directly into the data loading process, ensuring that unnecessary data is excluded from memory. This approach effectively manages memory constraints, particularly during the Filtration stage.

2) **Merging Data**: After loading and filtering the relevant data, the next step involves merging the different log files based on their timestamps. By combining logs from various sources such as logon, device, file, email, and HTTP logs, a unified sequence of events for each user is created. This sequence reflects the user's activities over time, capturing the temporal relationships between different actions.
3) **Session-Based Sequence Extraction**: Once the logs are merged, session-based sequences are generated by tracking logon and logoff events. Each session represents a continuous period of user activity, bounded by a logon at the start and a logoff at the end. All actions performed by a user between these events are grouped into a single session. This step is essential for understanding user behavior within specific time frames, which is crucial for detecting potential insider threats.
4) **Encoding**: To prepare the data for training the model, categorical variables must be transformed into numerical representations. This process includes several sub-steps:

    - **Feature Encoding**: Categorical variables are encoded into numerical form, ensuring that the model can process these variables effectively.
    - **Feature Scaling**: Numerical features are scaled to normalize their values, ensuring that all features contribute equally to the model training process.
    - **Tokenization and Padding**: Text sequences, particularly in the 'Activity' field, are converted into numerical sequences through tokenization. Padding is then applied to these sequences to ensure a consistent input length across all sessions, which is essential for feeding the sequences into the LSTM.

5.2. Data Splitting

The dataset was divided into two sets to facilitate model training and evaluation. 70% of the data was used for training, allowing the LSTM model to learn patterns associated with both normal and abnormal user behaviors. The remaining 30% of the data was allocated for testing, which was utilized to assess the model's performance in accurately detecting insider threats.

5.3. LSTM

It is a Long Short-Term Memory network (LSTM) which is a specialized form of Recurrent Neural Network (RNN) designed to model sequences, such as sequences of actions or languages. LSTM describes the sequence through a structure known as a memory cell, which contains three key gates: an input gate, a forget gate, and an output gate. These gates manage the flow of information, allowing LSTM to learn and train on long-term dependencies. The benefit of LSTM is that it maintains the order of user actions during those sessions.

The importance of LSTM lies highly in the effective processing, with the ability to automatically learn features. Since It has the ability to handle the sequence of actions in time series data it increases its strength making it ideal for modeling user behavior as required. The primary reason for utilizing LSTM is that the user behavior features this solution focuses on analyzing the temporal categorical features that occur at different timestamps. LSTM has the ability to remember long-term sequences of actions in sequential data making it suitable for this proposed solution. The second reason for choosing LSTM for this project is that it shows the likelihood of an individual user being an insider threat caused by their previous sequence of actions and behavior patterns.

5.4. Model Training

The LSTM model is constructed using a Sequential architecture specifically designed to capture and analyze the temporal dependencies inherent in user activity sequences, which is crucial for effective binary classification in insider threat detection. The model begins with a single LSTM layer comprising 100 units. This layer is responsible for processing the input sequences, representing user activities over time, and detecting patterns that may indicate normal or abnormal behavior. The temporal nature of LSTM allows the model to remember important information from earlier time steps while disregarding irrelevant details, making it well-suited for analyzing sequential data where the order of events is significant.

Following the LSTM layer, the model incorporates a Dense layer with a single neuron, which uses a sigmoid activation function. This layer is responsible for performing the binary classification task, outputting a probability that indicates whether a given session is likely to be abnormal or not. The sigmoid activation ensures that the output is a value between 0 and 1, which can be interpreted as the likelihood of an insider threat.

The model is trained over 100 epochs, which provides sufficient iterations for the model to learn from the data, while a batch size of 32 is used to balance the trade-off between computational efficiency and the granularity of updates to the model weights. Additionally, a validation split of 20% is employed during training, allowing the model's performance to be monitored on unseen data. This validation process helps in adjusting the model's parameters

and preventing overfitting by providing early insights into how well the model generalizes to new data.

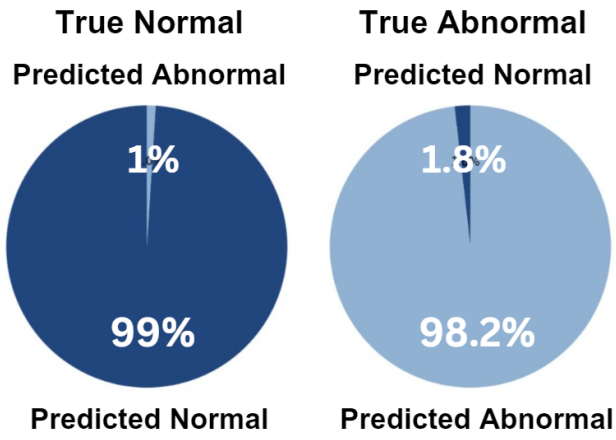


Figure 3. The Confusion matrix

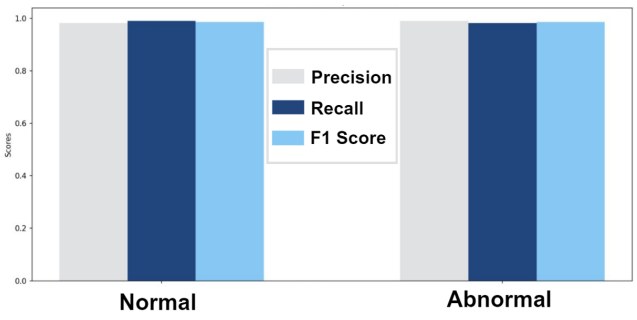


Figure 4. The Classification report matrix

5.5. Model Evaluation and Results

This section assesses the performance of the LSTM model used for insider threat detection, focusing on its ability to differentiate between normal and abnormal user activities. The evaluation process is tested by how the model is assessed on the test set, reporting both the loss and accuracy of its training. Then, the prediction part of the evaluation focuses on the model performance and how it generates predictions of normal and abnormal actions for the user activities, converts probabilities into binary labels, and evaluates its performance through a confusion matrix. Additionally, the classification report, which provides insights into the model's accuracy, precision, recall, and F1-score.

Fig. 3 displays the confusion matrices for the model's predictions, with results represented as percentages. The confusion matrices offer a clear view of the model's prediction accuracy for both normal and abnormal activities:

- For true abnormal activities, the model correctly identifies 98.2% as abnormal, with a small error margin of 1.8% where abnormal activities are misclassified as normal.
- For true normal activities, the model demonstrates 99% accuracy in predicting them as normal, with only 1% incorrectly classified as abnormal.

These results indicate that the model is highly reliable in distinguishing between normal and abnormal behaviors, achieving a strong true positive rate for both categories.

Fig. 4 presents the classification report matrix, summarizing the model's overall performance metrics. The report shows:

- Accuracy: 99%
- Precision: 99% for both normal (label 0.0) and abnormal (label 1.0) activities
- Recall: 99% for both normal (label 0.0) and abnormal (label 1.0) activities
- F1-Score: 99% for both normal (label 0.0) and abnormal (label 1.0) activities

The macro and weighted averages also reflect this high level of performance, underscoring the model's robustness and effectiveness in identifying insider threats with high accuracy.

6. Conclusion

The rise of insider threats causes significant challenges and issues to organizational security and privacy, requiring them to implement effective detection methods and solutions. This paper presented InternalEYE, a machine learning-based approach utilizing Long Short-Term Memory (LSTM) model to identify anomalies in user behavior. By analyzing patterns over time, InternalEYE enhances the detection of potential insider threats, achieving high accuracy and significantly reducing false positives compared to traditional security measures.

This solution not only helps making a safer workplace environment but also provides a proactive security culture within organizations. The model's high accuracy in differentiating between normal and abnormal behaviors minimizes errors, ensuring insiders threats are effectively detected. Furthermore, this solution provides a high early detection rate allows organizations to respond swiftly and quickly to potential risks, reducing the impact and potential damage of security incidents. The advanced anomaly detection capabilities empower organization to proactively address suspicious activities by employees, safeguarding sensitive data and maintaining organizational integrity. As insider threats become increasingly sophisticated, the integration of advanced analytical tools like InternalEYE is essential for fostering a secure and resilient operational landscape.

7. Future Work

The focus will be on developing a real-time detection system that includes a fully functional tool featuring an intuitive dashboard and an alert system. This system will enable continuous monitoring of insider threats and facilitate swift responses to any suspicious activities.

The intuitive dashboard will provide security teams with a clear and organized view of potential threats, allowing for quick decision-making and action. Additionally, this solution will integrate seamlessly with existing security detection tools. This integration will enhance security event correlation across the organization, allowing for a more comprehensive analysis of security data and improving the overall effectiveness of threat detection.

By combining these features, the system will improve to a real-time detection system will empower organizations to not only identify insider threats more effectively but also respond proactively, ensuring the protection of sensitive information and maintaining a secure operational environment.

References

[1] "What are insider threats?," Ibm.com, 19-Jul-2021. [Online]. Available: https://www.ibm.com/topics/insider-threats. [Accessed: 30-Aug-2024].

[2] Press Release, "Insider cyberthreats: 27% of companies in Saudi Arabia suffered from malicious actions by staff," Zawya, 10-Jan-2024. [Online]. Available: https://www.zawya.com/en/press-release/research-and-studies/insider-cyberthreats-27-of-companies-in-saudi-arabia-suffered-from-malicious-actions-by-staff-h2noolqd. [Accessed: 30-Aug-2024].

[3] Melanie, "Dense neural networks: Understanding their structure and function," Data Science Courses — DataScientest, 05-Mar-2024. [Online]. Available: https://datascientest.com/en/dense-neural-networks-understanding-their-structure-and-function. [Accessed: 31-Aug-2024].

[4] B. Lindauer, "Insider Threat Test Dataset." Carnegie Mellon University, 30-Sep-2020.

[5] F. Yuan, Y. Cao, Y. Shang, Y. Liu, J. Tan, and B. Fang, "Insider threat detection with deep neural network," in Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 43–54.

[6] [A. Saaudi, Z. Al-Ibadi, Y. Tong, and C. Farkas, "Insider threats detection using CNN-LSTM model," in 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018.