Insider Threat Detection in Corporate Networks Using Machine Learning Models

Muntaka Mubarrat Antorik
Department of Computer Science
BRAC University
Dhaka, Bangladesh
ID: 24241061
muntaka.mubarrat.antorik@g.bracu.ac.bd

Sujit Kumar Datta
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
ID: 22201946
sujit.kumar.datta@g.bracu.ac.bd

Tafsir Ul Hasan
Department of Computer Science
BRAC University
Dhaka, Bangladesh
ID: 24241053
tafsirulhasan@g.bracu.ac.bd

***Abstract*—Insider threats are a serious concern for modern organizations because employees already have authorized access to internal systems, making malicious actions difficult to detect using traditional security tools. In this project, machine learning techniques are applied to identify insider threats by analyzing user behavior in a corporate network. The study is based on the CERT Insider Threat dataset, which contains simulated logs of employee activities such as email usage and system behavior. Multiple supervised machine learning models, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and AdaBoost, were trained and evaluated. Experimental results show that machine learning models—particularly ensemble-based approaches—achieve strong performance in distinguishing malicious insider activity from normal behavior. Overall, this project demonstrates that machine learning can play a valuable role in strengthening insider threat detection systems in organizational environments.**

***Index Terms*—Insider Threat Detection, Machine Learning, Cybersecurity, CERT Dataset, Ensemble Models**

I. Introduction

With the exponential growth of digital infrastructure and organizational communication networks, today's organizations rely more on internal networks, emailing services, and information systems to facilitate their day-to-day operations. Though much has been done to safeguard organizations from external cyber threats, it is generally agreed today that the toughest and most perilous security threats to organizations today are indeed from internal sources or insiders. Compared to external hackers who lack internal access to organizational systems and resources, insiders already have authorized access to sensitive resources within organizations.

Insider threats can come with a variety of motivations, ranging from monetary gain, personnel grievances, complacency, or accidental errors. Such threats can lead to severe consequences such as data exfiltration, intellectual property theft, system tampering, and reputation damage. Classical security methods, such as rule-based monitoring or signature-based detection systems, are largely inefficient in the detection of insider threats because they are dependent on predefined rules or known attack signatures. Insider behavior tends to be more subtle, creeping, and context-dependent in nature.

To address the limitations presented by the traditional approaches, machine learning (ML) has proven to be an effective solution for insider threat detection. Using ML, complex patterns present in past data can be learned, allowing for the detection of malicious behavior that does not fall into the normal usage patterns of users. Using large data amounts present in activity logs in an organization, insider threats can be detected by an ML system with a larger accuracy rate than traditional approaches.

In the proposed project, insider threat identification will be discussed in terms of a machine learning classification issue using the CERT Insider Threat Dataset, which is a popular benchmark dataset that preserves logs of corporate activity with a high degree of realism. The CERT Insider Threat Dataset preserves logs of corporate activity with a high degree of realism and captures a high level of detail in their email communications with both structured data and free text. This data captures realistic working scenarios, forming an excellent basis for modeling insiders.

The key objective of this research is to perform the analysis of user email behavior in the corporate network and, accordingly, classify such behavior as normal or potentially malicious. In this context, extensive data preprocessing and feature engineering are done to transform raw email logs into meaningful numerical representations. Missing value handling, encoding categorical features, and deriving behavioral metrics will be used to capture attributes representative of communication frequency, attachment usage, and interaction patterns.

High-dimensional data often introduces noise or may lead to computational inefficiency. Dimensionality is, therefore, reduced with Principal Component Analysis (PCA) while retaining relevant behavioral information.

Besides supervised learning, this project also develops unsupervised behavioral grouping using K-Means clustering. Clustering enables users to be grouped based on similarities in activity patterns that provide insight into natural behavior segments and help visualize distinctions among normal versus suspicious users. This step is supportive of exploratory analysis and enhances the understanding of underlying data structure before classification.

In this paper, several machine learning algorithms are implemented and compared for supervised learning to assess the efficiency of insider threat detection. The algorithms used in this work include Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, KNN, Naive Bayes, and AdaBoost. Both simple and ensemble-based models are used to enable a comparison between linear, nonlinear, and ensemble learning. Accuracy and cross-validation are used to assess the performance of the models for their reliability and generalization on unseen data.

By applying this systematic approach, the project will be able to answer questions like these:

How well do machine learning algorithms classify normal and malicious insider actions? Can ensemble models outperform basic classifiers in insider threat identification tasks? The effectiveness of methods for reducing dimension and clustering in knowing patterns in insider behavior.

In summary, this project shows the application of various machine learning algorithms on a real-world cybersecurity issue, namely insider threat detection within the enterprise setting. The use of multiple phases, including data processing, dimensionality reduction, and unsupervised and supervised learning, shows the viability of developing ML-driven systems for the improvement of enterprise security. The result of this research helps understand the strengths and weaknesses of the various ML algorithms, setting the stage for the development of improvements, including temporal and real-time systems.

II. Literature Review

Recent studies on insider threat detection have focused on improving detection performance under severe class imbalance and limited availability of real-world attack data. One approach introduced a hybrid deep learning framework combined with a GAN-based data augmentation technique to generate realistic synthetic insider threat samples [4]. This method effectively addressed class imbalance and improved model generalization. However, the overall performance was shown to be highly dependent on the quality and representativeness of the original dataset.

Another significant research direction emphasized automated feature engineering to handle the heterogeneous and high-dimensional nature of insider activity logs [2]. Techniques such as Deep Feature Synthesis were used to automatically extract behavioral features from raw system logs, followed by dimensionality reduction using Principal Component Analysis and data balancing using SMOTE. Experimental results on the CERT dataset demonstrated that effective feature engineering substantially improves detection accuracy when combined with supervised learning models.

Several studies specifically investigated the impact of imbalanced data on insider threat detection performance [2]. Data-level solutions such as oversampling and undersampling were applied prior to model training, and ensemble learning methods were found to outperform individual classifiers. Despite achieving high accuracy, these approaches raised concerns regarding real-world applicability due to the reliance on synthetic data generation.

Supervised machine learning techniques have been shown to be more effective than traditional anomaly detection methods, which often suffer from high false-positive rates [1]. Ensemble-based models such as Random Forest and XGBoost consistently demonstrated superior performance in detecting insider threats. However, their effectiveness remained sensitive to class imbalance, highlighting the need for robust data preprocessing and evaluation strategies.

Comprehensive surveys of recent advancements in insider threat detection highlighted the growing adoption of machine learning, deep learning, NLP-based, and graph-based approaches [3]. These studies identified persistent challenges including data imbalance, privacy concerns, limited labeled datasets, false alerts, and lack of model explainability, indicating important directions for future research.

Recent work has also explored real-time insider threat detection systems using fast and scalable machine learning models [5]. These systems incorporated dynamic risk profiling to continuously assess user behavior and enable proactive threat mitigation. While promising, many real-time approaches relied on synthetic datasets due to the scarcity of real-world insider threat data.

Hybrid detection frameworks combining unsupervised and supervised learning have been proposed to improve efficiency under extreme class imbalance [6]. Unsupervised outlier detection techniques were used to highlight suspicious behavior, which was then refined using supervised classifiers. Although these methods reduced computational cost, their reliance on batch processing limited real-time responsiveness.

Graph-based and sequence-aware models have recently gained attention for capturing fine-grained activity-level behavior [7]. By learning adaptive relationships among user activities, these models achieved strong performance even under highly imbalanced conditions. However, their effectiveness depended heavily on the availability of high-quality and consistent activity logs.

Overall, existing literature demonstrates significant progress in machine learning-based insider threat detection, while challenges related to data imbalance, real-time deployment, explainability, and data quality remain open research problems.

III. Dataset Description

The dataset used in this project is the email.csv file from the CERT Insider Threat collection, which contains approximately 266,000 email-level records. Each record includes attributes such as id, date, user, pc, sender and recipient information, message size, attachment count, and email content. These logs span several months, allowing analysis of user communication behavior over time. To make the dataset suitable for machine learning and computationally efficient, raw email-level records were transformed into higher-level behavioral features. Multiple email events were aggregated into summarized behavioral profiles capturing communication frequency, attachment usage, and interaction patterns.

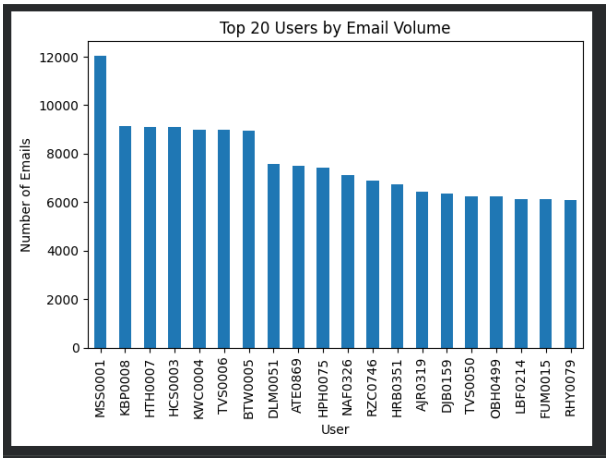

Fig. 1. Top 20 users by email volume in the CERT Insider Threat dataset, showing highly active users within the corporate network.

Figure 1 illustrates the distribution of email activity across users, highlighting a small group of highly active users and motivating user-level aggregation during feature engineering.

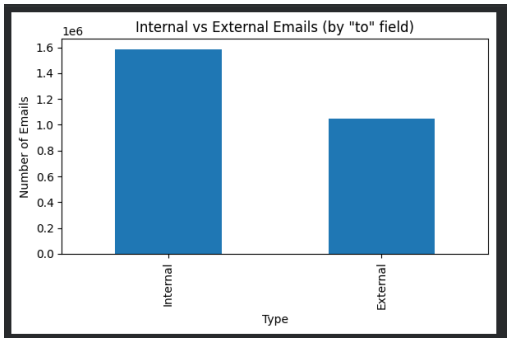

Fig. 2. Comparison of internal and external email communications based on recipient domains in the CERT dataset.

As shown in Figure 2, internal communications dominate the dataset, while a significant portion of emails are sent externally, which is a critical indicator for insider threat analysis.

After pre processing and feature engineering, a random sampling strategy was applied to select 1,000 representative instances while preserving the overall distribution of normal and suspicious user activities. The final dataset consists of 1,000 samples with 13 engineered features, enabling efficient model training while retaining meaningful behavioral information.

Fig. 3. Statistical summary of the CERT Insider Threat dataset showing feature distributions, counts, and descriptive statistics.

IV. Methodology

A. Data Preprocessing

The CERT Insider Threat dataset was preprocessed to ensure data consistency and suitability for machine learning analysis. Missing values and incomplete records were removed, and timestamp attributes were transformed into numerical and temporal features to capture user activity patterns.

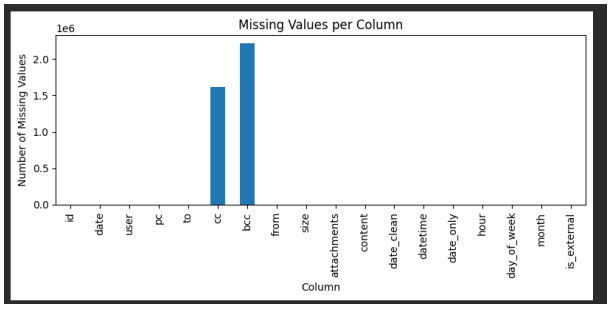

Fig. 4. Number of missing values per feature in the CERT Insider Threat dataset prior to preprocessing.

Figure 4 shows that fields such as *cc* and *bcc* contain a large number of missing values, which motivated their exclusion or transformation during preprocessing.

Categorical variables, including user identifiers and system-related attributes, were encoded using label encoding techniques. Feature engineering was applied to extract behavioral metrics such as email activity frequency, attachment usage, and communication patterns. To improve model efficiency and reduce noise, Principal Component Analysis (PCA) was employed for dimensionality reduction, eliminating redundant features while preserving the most informative components.

B. Model Development

Following data preprocessing and feature engineering, the insider threat detection problem was formulated as a binary classification task, where user activities were classified as either normal or malicious. All extracted features were first standardized to ensure uniform scaling across dimensions. Principal Component Analysis (PCA) was then applied to reduce feature dimensionality and remove noise, improving model efficiency and generalization. The PCA-transformed feature set was used consistently across all models to ensure a fair and uniform comparison.

In addition to supervised learning, an unsupervised clustering technique was applied to the reduced feature space to analyze behavioral patterns and visually examine the separation between normal and malicious activities. K-Means clustering was employed for exploratory analysis and visualization purposes, while supervised classification models were trained and evaluated using accuracy and cross-validation metrics.

1) Machine Learning Models Used: **Logistic Regression:** Logistic Regression was trained on the PCA-reduced dataset and used as a baseline classifier. The model demonstrated strong performance in distinguishing between normal and suspicious user behavior, indicating that the transformed feature space provided effective linear separability.

Decision Tree: A Decision Tree model was applied to capture non-linear decision boundaries present in insider behavior data. Although it achieved reasonable performance, its accuracy was slightly lower than ensemble-based methods, likely due to its sensitivity to noise and overfitting.

Random Forest: Random Forest was implemented as an ensemble of decision trees trained on the PCA-transformed features. The model achieved high accuracy and exhibited improved stability and generalization compared to a single decision tree.

Gradient Boosting: Gradient Boosting was used to iteratively improve classification performance by emphasizing previously misclassified instances. The model performed consistently well in detecting suspicious insider activities and ranked among the top-performing classifiers.

Support Vector Machine (SVM): SVM was trained on the reduced feature space and effectively separated normal and malicious behavior patterns. The application of PCA reduced computational complexity and contributed to high classification accuracy.

K-Nearest Neighbors (KNN): KNN was implemented using the PCA-reduced and standardized features. While the model was able to classify insider behavior reasonably well, its performance was influenced by data distribution and class imbalance.

Naive Bayes: Naive Bayes was employed as a probabilistic classifier on the transformed dataset. Despite its strong independence assumptions, it achieved competitive results with low computational overhead.

AdaBoost: AdaBoost was applied as a boosting-based ensemble technique that focuses on misclassified samples during training. The model achieved competitive accuracy and improved detection of difficult insider behavior patterns.

2) Unsupervised Model: **Clustering of K-Means:** K-Means clustering was applied to the PCA-reduced feature space to group users based on behavioral similarity. The clustering results were primarily used for exploratory analysis and visualization of insider behavior patterns rather than direct classification.

C. Tools Used

All experiments were conducted using Python in a Jupyter Notebook environment. Data preprocessing and feature engineering were performed using Pandas and NumPy, while model development, dimensionality reduction, clustering, and evaluation were carried out using Scikit-learn. Visualization of feature distributions and model performance was achieved using Matplotlib, enabling clear analysis, reproducibility, and efficient experimentation throughout the project.

V. Results and Evaluation

The proposed insider threat detection system demonstrated strong performance across all evaluated machine learning models after preprocessing and dimensionality reduction using PCA. Cross-validation results show that Logistic Regression achieved the highest accuracy (98.29%), followed by Support Vector Machine (97.57%), Random Forest (97.43%), and Gradient Boosting (97.14%). Naive Bayes, Decision Tree, and KNN also produced competitive results with slightly lower accuracy values. In addition to accuracy, models were evaluated using precision, recall, F1-score, and confusion matrices to provide a comprehensive assessment. Classification reports indicate that ensemble-based models achieved strong precision and recall, demonstrating their effectiveness in correctly identifying malicious insider behavior while minimizing false alarms. Confusion matrix analysis further confirms high true positive and true negative rates across most models, indicating reliable separation between normal and suspicious user activities. These results highlight the robustness of the engineered features and the effectiveness of the proposed machine learning framework.

Figure 5 compares the classification accuracy of the supervised machine learning models used in this study.

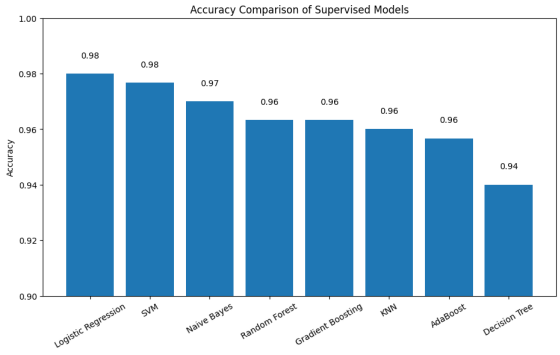

Fig. 5. Accuracy comparison of supervised machine learning models used for insider threat detection.

To better understand the separation between normal and suspicious behaviors, PCA was applied to reduce feature dimensionality while preserving important behavioral patterns. Figure 6 presents a visualization of insider behavior using the first two principal components. The plot demonstrates a noticeable separation between normal and suspicious user activities, indicating that the extracted features are effective in capturing behavioral differences.

Figure 7 shows the cross-validation accuracy distribution across the evaluated machine learning models. The relatively

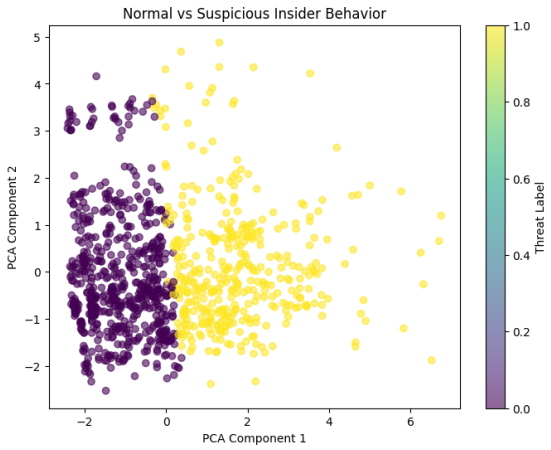

Fig. 6. Visualization of normal and suspicious insider behavior using the first two principal components obtained from PCA.

balanced distribution and high accuracy values across models suggest that the engineered behavioral features are robust and generalize well across different classifiers, further validating the effectiveness of the proposed approach.

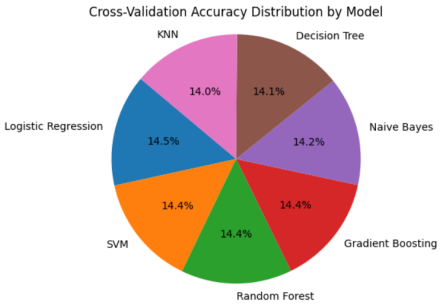

Fig. 7. Cross-validation accuracy distribution of machine learning models used for insider threat detection.

VI. Discussion

The experimental results demonstrate that the proposed machine learning approach is effective for insider threat detection after comprehensive preprocessing and dimensionality reduction using Principal Component Analysis (PCA). Ensemble-based models such as Random Forest and Gradient Boosting achieved consistently high accuracy by capturing complex and non-linear behavioral patterns in user activity data. At the same time, simpler models, particularly Logistic Regression, also performed competitively on the PCA-reduced feature set, indicating that the engineered features provided meaningful separation between normal and suspicious user behavior. The strong performance across multiple classifiers suggests that the modeling pipeline is robust and not dependent on a single algorithm. PCA played a significant role in reducing feature redundancy and noise, enabling both linear and non-linear models to generalize effectively. Support Vector Machine and AdaBoost also showed stable performance, further validating the effectiveness of the selected features. K-Means clustering was used as an exploratory technique to analyze user behavior patterns in the reduced feature space. The clustering results revealed distinguishable groupings of users, supporting the assumption that insider activity differs measurably from normal behavior, although clustering was not directly used for prediction. Despite these positive outcomes, several limitations remain. The CERT dataset is simulated and may not fully capture real-world insider behavior. Additionally, temporal dependencies and sequential activity patterns were not explicitly modeled, which may limit the detection of long-term or evolving threats. The presence of class imbalance further suggests that future evaluations should include additional metrics beyond accuracy. Future work may focus on incorporating temporal and deep learning models, such as recurrent or transformer-based architectures, as well as extending the framework to real-time detection using streaming data. Validation on real-world datasets and improved evaluation strategies would further enhance the applicability of the proposed approach.

VII. Conclusion

This project investigated the application of machine learning techniques for detecting insider threats within corporate environments using the CERT Insider Threat dataset. Through comprehensive data preprocessing, feature engineering, dimensionality reduction using Principal Component Analysis (PCA), and behavioral grouping with K-Means clustering, multiple supervised classification models were trained and evaluated.

The results indicate that ensemble-based models, particularly Logistic Regression and SVM, achieved strong performance in distinguishing malicious insider activity from normal user behavior. These findings highlight the potential of machine learning to enhance insider threat detection systems by identifying complex behavioral patterns that traditional rule-based approaches often overlook. With further refinement, incorporation of temporal modeling, and validation on real-world datasets, such machine learning-driven approaches could be effectively deployed in practical cybersecurity environments.

References

[1] P. Manoharan et al., "Insider threat detection using supervised machine learning algorithms," *Telecommunication Systems*, vol. 87, no. 4, pp. 899–915, 2023.

[2] B. B. Sarhan and N. Altwaijry, "Insider threat detection using machine learning approach," *Applied Sciences*, vol. 13, no. 1, p. 259, 2022.

[3] F. R. Alzaabi and A. Mehmood, "A review of recent advances in malicious insider threat detection," *IEEE Access*, vol. 12, pp. 192827–192840, 2024.

[4] R. G. Gayathri et al., "Hybrid deep learning model using SPCAGAN augmentation for insider threat analysis," arXiv:2203.02855, 2022.

[5] S. A. Qawasmeh and A. A. S. AlQahtani, "Beyond firewall: Leveraging machine learning for real-time insider threats identification," Preprints, 2025.

[6] J. Yi and Y. Tian, "Insider threat detection model enhancement using hybrid algorithms," *Electronics*, vol. 13, no. 5, p. 973, 2024.

[7] X. Cai et al., "Learning adaptive neighbors for real-time insider threat detection," arXiv:2403.09209, 2024.