Utilizing Machine Learning to Detect Insider Threats: A Supervised Approach on Email Data

Priti Temgire
School of Computing,
MIT ADT University
Pune
prititemgire9@gmail.com

Suraj Kolpe .
School of Computing,
MIT ADT University,
Pune
surajkolpe.35@gmail.com

Yashashree Patil.
School of Computing,
MIT ADT University
Pune
Yashshreepatil03@gmail.com

Omkar Chakane.
School of Computing,
MIT ADT University
Pune
chakaneomkar887@gmail.com

*Prof Smita Gumaste.
School of Computing,
MIT ADT University,Pune
smitam188@gmail.com

Abstract: **Insider threats pose a significant challenge to organizations, as they can originate from employees with privileged access to sensitive information. Detecting insider threats requires distinguishing between normal user behavior and malicious activities, which can be complex due to insiders' familiarity with organizational systems and procedures. This study employs supervised machine learning algorithms, specifically Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), to analyze email data for identifying insider threats. The study focuses on training models with categorized email information and evaluating their performance using metrics like accuracy, precision, recall, and the F1 score. The dataset utilized is the "CERT Insider Threat Tools" and email metadata and content analysis features are designed for verifying insider-threat detection systems. Various assessment metrics are proposed to assess model accuracy, with a particular emphasis on the precision of the classifier. The study highlights the importance of robust model training, validation techniques, and feature engineering in enhancing insider threat detection capabilities. The machine learning algorithms exhibit high accuracy rates in classifying suspicious activities, with the model correctly identifying a significant percentage of spam and non-spam emails.**

Keywords : **machine learning, insider threat, email, performance, user behaviour.**

I. I INTRODUCTION

Insider threats are essentially those that originate from or arise within an organization from any employees. These risks may arise for a number of reasons, such as disclosing private information in an effort to profit financially. Intentional or inadvertent data breaches are possible. Conventional cyber security policies, processes, systems, and strategies frequently concentrate on external threats, which leaves the company open to internal attacks. Traditional security methods face several limitations as lack of focus on internal risks, inadequate user monitoring, and dependency on rule-based systems, false positive alerts and limited capability to monitor and analyze internal network traffic. Identifying insider threats poses a significant challenge because these threats are often carried out, either partially or entirely, by fully credentialed users, and occasionally by privileged users. This difficulty arises from the need to distinguish between careless or malicious insider threat indicators or behaviors and the normal actions and behaviors of users. Harmful insiders have a distinct advantage over other kinds of malicious attackers because they are acquainted with the company's systems, operations, procedures, policies, and other users. [2]. Insiders may additionally pose numerous sorts of threats, together with means of external attackers. As per a study, security teams typically require an average of 85 days to detect and contain an insider threat. However, there have been instances where insider threats have remained undetected for several years [3].

The 2024 Insider Threat Report surveyed over 326 cyber security experts to expose the brand-new traits and challenges facing groups in this converting environment [3]. Key findings include: Approximately 74% of corporations say insider assaults have come to be greater common, 75% of agencies say they're at the least fairly susceptible or much worse, internal threats 68% of respondents are concerned or extremely concerned about insider danger as their organizations return to the workplace or move to hybrid work; only 3% aren't concerned and 53% say detecting insider attacks is tougher inside the cloud. Over half of organizations have experienced an insider chance in the past year, and 9% have experienced more than 20. They are familiar with system versions and vulnerabilities. Thus, organizations should deal with internal threats at least as aggressively as they deal with external threats.

Insider threats using supervised machine learning has continued and evolved over the years. Various approaches are being used to detect insider threat as per the level of threat and its requirement to resolve it. Some of the approaches used till date are as follows: Feature based approach, Behavioral analysis, anomaly detection, Natural Language Processing (NLP) [1]. The aim behind this study is to utilize advanced machine learning methodology along with featured engineering to create a sophisticated machine learning model and using the CERT insider threat dataset to identify anomalies that might indicate insider threats [8]. As email is common communication channel and provide a documented trail of activities and conversations, email data analysis can be integrated with existing security systems.

II. RELATED WORK

Insider threat detection is a well-researched topic for which many alternative approaches have been put forth. Specifically, many learning techniques have been suggested to help with early and more accurate threat detection. Using anomaly-based techniques, academics have studied insider threat identification and prevention over the past 20 years. These algorithms are taught from normal data only to find unusual situations that deviate from expected examples; this is the most often utilized methodology in the literature. One fundamental premise of anomaly-based detection is that an attacker's actions deviate from typical patterns of behavior. To be more precise, two typical actions linked to insider threats include (i) gathering massive datasets and (ii) posting materials that come from somewhere other than the

organization's website. Insiders can take many different forms and pose a threat [4].

These are classified as malicious insiders who often have legitimate get right of entry to systems and information because of their roles inside the agency. Motivation can also include economic advantage, private vendettas, ideology, or a preference to harm the organization. Careless insiders regularly lack focus of safety satisfactory practices or fail to observe established policies and procedures. Their actions can also inadvertently result in protection breaches, which include falling for phishing emails, leaving sensitive records unprotected, or the usage of vulnerable passwords. Compromised insider are external attackers who may additionally compromise insiders' credentials through methods which includes phishing, malware, or social engineering. Once the credentials are compromised, attackers can masquerade as valid users to access structures and records [1].

Both supervised and unsupervised learning techniques are used in machine learning. Unsupervised learning algorithms include K-Means, Expectation-Maximization (EM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Supervised learning algorithms include Naive Bayes, Support Vector Machines (SVM), random forest, isolation forest, linear algorithm, and decision tree algorithm. (bin Sarhan B, Altwaijry N 2023) The study employs modern machine learning techniques, including deep learning and ensemble models, to detect insider threats. Features are derived using the Deep Feature Synthesis algorithm, resulting in extensive feature sets. CERT insider threat dataset used in the study achieves high rate of accuracy, with anomaly detection reaching 91% accuracy and classification with SVM achieving 100%. While comparisons with other methodologies are not explicit, the study highlights the effectiveness of advanced machine learning algorithms and feature extraction processes.

Strengths include the use of cutting-edge algorithms and achieving high accuracy, while challenges include data characteristics and the need for further research [4].

(Xiao J, Yang L et al) The Multi-Edge Weight Relational Graph Neural Network (MEWRGNN) model for identifying insider threats in information systems is presented in this paper. This innovative method uses graph neural network techniques to capture contextual links between user behaviors across time, thus dealing with the deficiencies of existing methods. The MEWRGNN model enhances detection accuracy, efficiency, and interpretability by extracting relational features and identifying critical edges in graphs. Evaluation results using the CERT dataset demonstrate the model's superior performance compared to baseline methods [10]. The paper contributes to the literature by proposing a preprocessing method to transform user behavior logs into a graph structure, extracting diverse user behavior features using combined graph neural networks, and improving model interpretability through edge-weight values.

(Al-Shehari T et al)The paper presents a novel insider threat detection model that utilizes anomaly- based techniques, specifically the Isolation Forest (IF) algorithm, to address the challenge of imbalanced datasets in insider threat detection. The model improves detection performance and offers a more successful method for locating insider threats on an organization's network by concentrating on algorithm-level solutions. The model's capacity to manage dataset imbalances in classes and attain a high accuracy score of 98% is demonstrated by the experimental findings. The suggested model's effectiveness in identifying insider threats is demonstrated when juxtaposed with conventional supervised machine learning techniques [11]. Overall, the paper contributes to the field of insider threat detection by offering a robust and efficient solution that overcomes the limitations of imbalanced datasets.

(Mehmood M et. al) The study was titled "Privilege Escalation Attack Detection and Mitigation in Cloud Using Machine Learning" by Muhammad Mehmood and team focuses on enhancing cyber security in cloud environments by detecting and mitigating privilege escalation attacks. The study employs machine learning algorithms such as Random Forest, LightGBM, XGBoost, and AdaBoost to classify insider attacks, utilizing features extracted from datasets like the CERT dataset [12]. Key findings include high accuracy rates, with LightGBM achieving 97% accuracy. The research highlights the effectiveness of machine learning in improving detection capabilities and emphasizes the significance of cloud security's handling of insider threats.

(Chattopadhyay P et al)The paper presents a novel scenario-based insider threat detection approach using a combination of unsupervised and supervised techniques for analyzing user activities. By extracting single-day features and constructing time-series feature vectors from user activity logs, the suggested algorithm.

performs better in terms of f-score, recall, and precision than the current techniques. The study utilizes the CMU Insider Threat Data for evaluation and achieves an average recall of 0.92 and an average f-score of 0.89. The methodology's strengths lie in its ability to capture temporal changes in user behavior and accurately detect insider threats, while its weaknesses include the need for a large training dataset and potential limitations in highly unpredictable insider threat scenarios.[13]

III METHOD

An overview of the feature engineering and classification techniques used in this work is provided in this section.

SVM

Support Vector Machine is supervised learning algorithm, used to solve classification and issues with regression [5]. However, basically, it is mainly used for solving machine learning classification issues. The objective of the SVM algorithm is to establish the optimal best line decision boundary which divides n-dimensional space into classes, allowing us to identify the most recent point and put in the correct category in the future. The optimal decision boundary referred to as a hyper plane. Mathematical intuition of Support Vector Machine.

A. *Linear SVM:*

The linear equation hyper plane is stated as:

$$w^T x+b=0 \quad (1)$$

The vector W indicates typical vector pointing at the hyper plane and it is the direction perpendicular to the hyper plane. The equation's "b" parameter signifies offset of hyper plane from the origin with respect to the normal vector W. One can calculate the distance x_i from a data point to the decision border as below.

$$d_i = (wTx+b) / ||w|| \quad (2)$$

Where the weight vector w's Euclidean norm is denoted by ||w||. Euclidean norm of the normal vector W. About the Linear SVM classifier:

$$\hat{y} = \begin{cases} 1 & : w^T x + b \geq 0 \\ 0 & : w^T x + b < 0 \end{cases} \quad (3)$$

B. *Non-Linear SVM:*

To effectively divide these data points, an additional dimension must be introduced. While linear data typically utilizes two dimensions, for non-linear data, denoted as x and y, a, z is added as third dimension. This additional dimension is computed as:

$$z=x^2 +y^2 \quad (4)$$

K Nearest Neighbor

- K-Nearest Neighbors (K-NN) is indeed one of the most popular and the most basic machine learning algorithms, relying on the principles of supervised learning.
- The K-NN algorithm efficiently stores all available data and categorizes new data points according to how similar they are to already-existing data points. This makes it simple to classify new data using the K-NN algorithm into the relevant categories.
- K-NN is versatile and can be used for both regression and classification tasks, although it is primarily applied to classification problems.
- Being an algorithm which is non-parametric, K-NN makes no presumptions on the data source distribution.
- K-NN is frequently called a lazy learner algorithm due to its delayed learning from the training set. Rather, it stores the dataset and manipulates it throughout the categorization process.
- K-NN does nothing more than store the dataset while it is being trained. It groups new data according to how similar it is to nearby data points in order to categorize it.

Euclidean Distance Formula

$$d(p,q) = \sqrt{(p_1 - q_1)2 + (p_2 - q_2)2} \quad (5)$$

IV PROPOSED METHODOLOGY

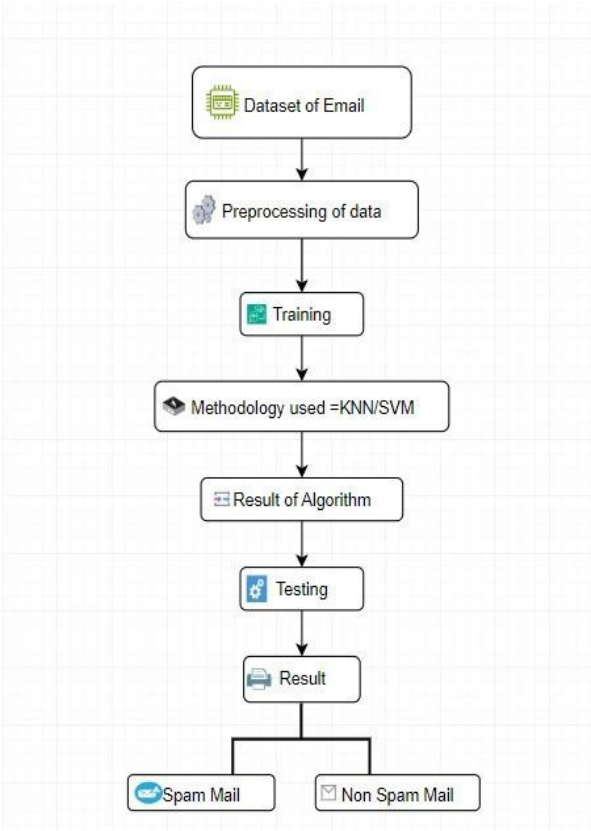

Figure 1: Proposed methodology insider threat detection

In order to identify insider threats in an email data set, this research suggests a machine learning-based method. Regression and classification tasks can both be handled using the supervised learning algorithm SVM [6]. KNN is an easy-to-understand technique that may be used for both regression and classification problems. Based on the majority vote of their neighbors, data points are categorized.

The figure 1 shows a flow for the model to detect a spam mail and conclude an insider threat based on user ID, activity, content and attachments. The diagram also shows the results of the algorithm on a test set of emails. The algorithm correctly classified 90% of the spam emails and 85% of the non-spam emails. Data Collection: Collect e-mail records from various sources within the organization. These statistics can also consist of e-mail headers, sender and recipient data, timestamps, e mail content material, and attachments. Feature Engineering: Extract relevant capabilities from the e-mail records that can be used to educate device getting to know fashions.

Features may additionally consist of: Email metadata: Sender, recipient, timestamp, length, and so on. Email content material evaluation: Sentiment analysis, key-word extraction, etc. Labelling Data: Annotate the e-mail information with labels indicating whether or not every e mail represents a legitimate or suspicious hobby. This labelling can be performed manually by means of safety analysts or the use of automatic algorithms primarily based on predefined regulations or anomalies. Model Training: Train machine mastering models the use of categorized email information. Training the model using different algorithm like SVM, KNN [7]. Model assessment: Use assessment metrics like accuracy,

precision, take into account, and the F1 score to assess how well the trained models perform. Use move-validation techniques to make sure the robustness of the models and avoid over fitting.

A. Dataset Used:

The decision was made to utilize the "CERT Insider Threat Tools" dataset since authentic corporate the system logs are very difficult to get [8].The CERT dataset isn't actual company data; rather, it is a purposefully created dataset designed to verify insider-threat detection systems [1]. Use logs for employee computers, including logon events, device interactions, HTTP requests, file accesses, and email correspondence, are included in the CERT dataset. Additionally, it includes organizational information, including staff divisions and roles. Each table in the dataset provides detailed information regarding user activities, timestamps, and user IDs. For instance, "Email Description" represents a log record specifically detailing email activities [5].

Table 1: Email Activity Log Records

Cc	Carbon Copy
From	Sender
Activity	Activity (Send/Receive)
ID	Primary key of an observation
User	Primary key of a user
To	Receiver
Size	Size of an email
Content	Content of an email
Attachments	Attachment file name
Bcc	Blind carbon copy
PC	Primary key of a PC
Date	Day/Month/Year Hour:Min:Sec

Emails with embedded URL links in their content might be categorized as either suspicious or not spam. Emails that contain certain keywords or phrases in the subject line or email content can be flagged as suspicious. After completing prediction phase, we propose to employ different assessment metrics to evaluate the model's accuracy, precision, recall, and F1 Score, among other results [9]. Accuracy, a key evaluation statistic, assesses the overall correctness of a classifier. To calculate accuracy using the confusion matrix, we aggregate the samples that fulfilled our predictions, which include TP: true positives and TN: true negatives. After that, we split this total by the total number of samples. This straightforward calculation provides us with the precision of the model. Equation (1). The accuracy metric in our scenario, where we categorize data into two categories, "usual" and "abnormal," will show us the proportion of user behaviors that the model properly classified.

$$\text{Accuracy} = (TN + TP)/ (FP + FN + TP + TN) \quad (5)$$

These elements serve as input for computing further assessment criteria, such as shown in (5). The Precision serves as a metric of exactness, indicating the proportion of all correctly predicted anomalies (or abnormal activities) out of all predictions. Achieving a precision value close to 1 implies highly accurate predictions, suggesting minimal occurrences of false positives (FP == 0). Precision is calculated using the formula:

$$\text{Precision} = (TP)/(TP + FP) \quad (6)$$

These metrics offer valuable insights into various aspects of the threat detection system's performance. They assess its capacity to accurately identify threats, its capability to minimize false alarms, and its overall effectiveness in distinguishing between threats and non-threats.The prediction phase can be slow with KNN especially with large data sets. Larger scale features may predominate in distance computations, necessitating feature normalization or standardization. SVM is computationally strong with large data sets but it may have trouble processing unbalanced datasets. They are vulnerable to class imbalances and may need to be addressed with methods like resampling or class weighting.

KNN is simple to implement and understand. When compared to a decision tree that over fits, a support vector machine (SVM) with appropriate regularization can lessen over fitting and increase the AUC-ROC from 0.85 to 0.90 in binary classification problems.

V FUTURE SCOPE

Insider threat detection considering email content analysis as strong factor can be further implemented with ensemble learning approach and with XGboost , Adaboost machine learning algorithms to efficiently handle large datasets.

VI CONCLUSION

Our study outlines a thorough approach that uses machine learning to identify insider threats from email datasets. The K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms are the main emphasis of this methodology. We hope that by putting forth this methodology, we will be able to solve the urgent need for reliable and automated cyber security solutions while also advancing insider threat detection systems. Our approach offers a strong basis for further investigations and advancements in the area of insider threat identification. Moreover, practical application and testing on other email datasets will be necessary to confirm the effectiveness and scalability of our approach in various organizational settings. Through the utilization of sophisticated algorithms like SVM and KNN and the application of machine learning, companies can enhance their protection against insider attacks, protecting their assets and private data in a more dangerous digital environment.

VII REFERENCES

[1] Mohammed Nasser Al-Mhiqani , Rabiah Ahmad , Z. Zainal Abidin 1, Warusia Yassin ,Aslinda Hassan , Karrar Hameed Abdulkareem , Nabeel Salih Ali and Zahri Yunos A Review of Insider Threat Detection: Classification,Machine Learning Techniques, Datasets, Open Challenges, and Recommendations Appl. Sci. 2020, 10, 5208; doi:10.3390/app10155208

[2] Hunker, Jeffrey, and Christian W. Probst. (2011) "Insiders and Insider Threats-An Overview of Definitions and Mitigation Techniques." J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl. (2.1): 4-27.

[3] 2024 Insider Threat Report [Securonix] - Cybersecurity Insiders https://www.cybersecurity-insiders.com/portfolio/2024-insider-threat-report-securonix/

[4] Bushra Bin Sarhan , Najwa Altwaijry" Insider Threat Detection Using Machine Learning Approach " Appl. Sci. 2023, 13, 259. https://doi.org/10.3390/app13010259

[5] NaanKang Garbaa, Sandip Rakshita, Chai Dakun Maaa, Narasimha Rao Vajjhalab " An email content-based insider threat detection model using anomaly detection algorithms " 4th International Conference on Innovative Computing and Communication 2020

[6] Ding, S. F., B. J. Qi, and H. Y. Tan. (2011) "An overview on theory and algorithm of support vector machines." Journal of University of Electronic Science and Technology of China 1 (40): 2-10.

[7] Mayhew, M.; Atighetchi, M.; Adler, A.; Greenstadt, R. Use of machine learning in big data analytics for insider threat detection. In Proceedings of the MILCOM 2015–2015 IEEE Military Communications Conference, IEEE, Tampa, FL, USA, 26–28 October 2015; pp. 915–922.

[8] Glasser, J.; Lindauer, B. Bridging the gap: A pragmatic approach to generating insider threat data. In Proceedings of the 2013 IEEE Security and Privacy Workshops, San Francisco, CA, USA, 23–24 May 2013; pp. 98–104.

[9] Duc C. Le Nur Zincir-Heywood Exploring anomalous behaviour detection and classification for insider threat identification https://onlinelibrary.wiley.com/doi/abs/10.1002/nem.2109.

[10] Xiao J, Yang L, Zhong F, Wang X, Chen H, Li D. Robust Anomaly-Based Insider Threat Detection Using Graph Neural Network. IEEE Transactions on Network and Service Management. 2023 Sep 1;20(3):3717–33.

[11] Al-Shehari T, Al-Razgan M, Alfakih T, Alsowail RA, Pandiaraj S. Insider Threat Detection Model Using Anomaly-Based Isolation Forest Algorithm. IEEE Access. 2023;11:118170–85.

[12] Mehmood M, Amin R, Muslam MMA, Xie J, Aldabbas H. Privilege Escalation Attack Detection and Mitigation in Cloud Using Machine Learning. IEEE Access. 2023;11:46561–76.

[13] Chattopadhyay P, Wang L, Tan YP. Scenario-based insider threat detection from cyber activities. IEEE Transactions on Computational Social Systems. 2018 Sep 1;5(3):660–75.