International Journal of Advanced Trends in Multidisciplinary Research
ISSN: 3139-0986, Volume 2 Issue 7
July 2026, PP 1-6

Machine Learning-Based Insider Threat Detection Using CERT Dataset

T.Pravalika[1], B.Gayathri[2]
[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana, India.
[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana, India.

Abstract **– The fact that malevolent actions are carried out by persons with legal access to organisational resources makes insider threats one of the most difficult cybersecurity issues confronted by contemporary organisations. Since suspicious actions typically mimic legitimate user behaviour, insider threats are harder to identify than external intrusions. When dealing with very unbalanced datasets, where harmful events make only a tiny proportion of overall user actions, traditional rule-based security systems often fail to detect modest behavioural anomalies. In order to enhance organisational security via intelligent behavioural analysis, this research proposes a system for insider threat identification that is based on machine learning and uses the CERT 5.2 Insider Threat Dataset. A variety of machine learning algorithms, such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, AdaBoost, and XGBoost, are included in the suggested framework, along with thorough data preprocessing and the Synthetic Minority Oversampling Technique (SMOTE) to deal with class imbalance. To find out how well the models detect insider threats, we use Accuracy, Precision, Recall, and F1-score. The experimental findings show that among the machine learning models tested, Random Forest and AdaBoost surpass the rest with a classification accuracy of 97.5%, all while retaining outstanding recall and precision. While maintaining the original methodology, dataset, algorithms, and experimental results, the suggested framework offers a scalable and efficient solution for early insider threat detection. This helps organisations with cybersecurity, security risk reduction, and intelligent behavioural analysis for better decision-making.**

Keywords **– Insider Threat Detection, Machine Learning, CERT Dataset, SMOTE, Random Forest, AdaBoost, XGBoost, Cybersecurity, User Behavior Analysis, Organizational Security.**

I. INTRODUCTION

Organisations rely heavily on information systems, cloud computing, corporate apps, and networked infrastructures due to the fast expansion of digital transformation. The operational efficiency and communication have both been enhanced by these technological improvements, but they have also brought forth advanced cybersecurity threats. Since they come from people with permission to access company resources, insider threats are among the most serious security issues that may arise from these types of problems. Financial losses, operational interruption, intellectual property theft, and reputational harm may occur as a consequence of privileged users, contractors, employees, or partners compromising sensitive information, whether deliberately or accidentally. Insider threats are much more difficult to detect than external assaults since their actions often seem like normal business practices.

For the most part, traditional cybersecurity systems rely on signature databases, established security rules, and access control policies to detect malicious actions. These methods work well against common attack patterns, but they may miss insider threats because their behaviour is so nuanced. Due to ever-changing user behaviour and the massive amount of security-related data produced daily, insider threat detection is becoming more complicated in more complex organisational setups. Because of these restrictions, scientists are looking at machine learning methods that can learn patterns of behaviour and spot suspicious actions automatically, rather than depending just on security rules that have to be specified by humans.

Machine learning's ability to analyse massive datasets, uncover hidden behavioural links, and accurately categorise suspicious activity has made it a powerful tool in cybersecurity. Unfortunately, there is a class imbalance when it comes to insider threat detection, meaning that harmful insider behaviours only make up a tiny fraction of the data that is accessible. Predictions that favour typical user behaviour are often the result of training machine learning models on datasets that are skewed in this way. In order to circumvent this restriction and enhance the capacity of models to learn, balancing approaches like the Synthetic Minority Oversampling Technique (SMOTE) have been commonly used. Algorithms trained with machine learning are far better able to spot infrequent but crucial insider threat occurrences if we fix the class imbalance first.

Using the CERT 5.2 Insider Threat Dataset, this paper presents a methodology for insider threat identification based on machine learning. A number of machine learning classifiers, including as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, AdaBoost, and XGBoost, are assessed by the framework. It also applies SMOTE on the dataset in order to restore its balance. The best successful model for insider threat detection is determined by a comparison study utilising Accuracy,

International Journal of Advanced Trends in Multidisciplinary Research
ISSN: 3139-0986, Volume 2 Issue 7
July 2026, PP 1-6

Precision, Recall, and F1-score. The usefulness of Random Forest and AdaBoost in recognising harmful insider behaviour inside organisational contexts is highlighted by their greatest accuracy of 97.5%, as shown by experimental assessment.

Here are the main things this effort has contributed: Using the CERT 5.2 dataset, build a machine learning system for detecting insider threats. • SMOTE being used to efficiently handle a significant class imbalance. • Eight different machine learning algorithms were tested in a controlled environment to see how they compared. • Examination of results using F1-score, Accuracy, Precision, and Recall measures. • Determination that the best models for detecting insider threats in organisations are Random Forest and AdaBoost.

II. LITERATURE SURVEY

Research on intelligent algorithms that may detect hostile user behaviour inside organisational settings has been extensively pursued due to the rising incidence of insider assaults. Security mechanisms based on rules and intrusion detection systems powered by signatures were the primary areas of early research. Malicious insiders often have valid system access and work within conventional organisational bounds, making it difficult for these techniques to identify insider threats, even when they were effective in detecting established attack patterns. As a result, there has been a surge in the investigation of machine learning techniques that may learn behavioural traits and detect tiny irregularities linked to insider activity.

A number of scholars have reviewed insider threat detection approaches extensively. The increasing significance of intelligent behavioural analysis for organisational cybersecurity was emphasised by Sandra G. et al., who evaluated current advancements in insider threat identification. They discussed different datasets, detection methodologies, and ways for evaluating performance. Support Vector Machines (SVMs) and Recurrent Neural Networks (RNNs) were identified as two of the most popular methods for analysing user behaviour and network traffic patterns in Femi-Oyewole et al.'s analysis of deep learning and machine learning approaches to network-based insider threat detection.

The efficacy of ensemble learning algorithms in detecting insider threats has been further shown by comparative research. After comparing XGBoost against Random Forest, Support Vector Machine, K-Nearest Neighbours, Multi-Layer Perceptron, and Logistic Regression, among other machine learning classifiers, Abdallah et al. found that XGBoost had the best detection performance in all of the test cases. Another study by Sridevi et al. suggested a framework that combines deep learning with machine learning. This framework could detect subtle insider threat behaviours by analysing user activity logs and feature-engineered behavioural patterns. It outperformed conventional approaches in terms of prediction accuracy.

The significance of resolving class imbalance, a significant obstacle in insider threat identification, has been brought to light in recent studies. Because harmful actions only make up a tiny fraction of total organisational events, traditional machine learning algorithms tend to favour legitimate user actions. To enhance minority-class recognition while decreasing false positive rates, researchers have suggested oversampling approaches, anomaly detection methods, behavioural analytics, and ensemble learning procedures. On the whole, these studies show that insider threat identification is much improved when strong machine learning models are combined with balanced datasets.

In light of these results, the current study suggests a strategy for detecting insider threats using machine learning and the CERT 5.2 Insider Threat Dataset. Before analysing Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, AdaBoost, and XGBoost using Accuracy, Precision, Recall, and F1-score, the suggested method evenly distributes the dataset using SMOTE. Based on the results of the experiment, the top classifiers for realistic insider threat detection in organisational contexts are Random Forest and AdaBoost, which both achieve 97.5% accuracy. This proves that ensemble machine learning approaches are successful.

III. ML FRAMEWORK FOR INSIDER THREAT DETECTION

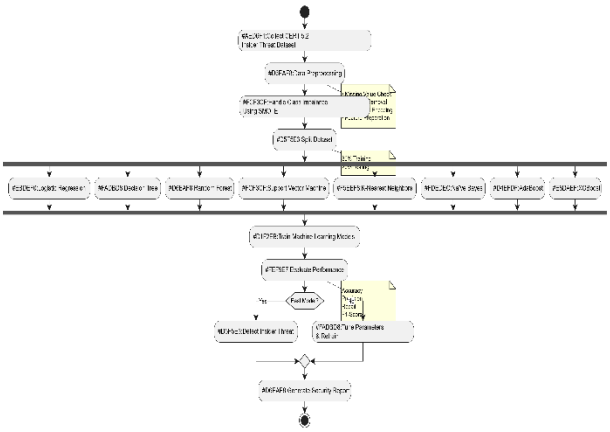

Fig. 1. Proposed Machine Learning Framework for Insider Threat Detection

Through the application of several machine learning algorithms to patterns in user behaviour and organisational operations, the suggested framework aims to detect insider threats. The suggested framework autonomously learns behavioural traits from organisational data history and detects suspicious activity using intelligent classification algorithms, in contrast to traditional security systems that mostly rely on preset rules or signature databases. An all-inclusive framework for detecting organisational insider threats is available; it includes data collecting, preprocessing, class balancing, feature extraction, training

of machine learning models, performance assessment, and threat prediction.

An industry standard for insider threat research, the CERT Insider Threat Dataset is the starting point for the detection procedure. Users' logins, file access operations, USB device usage, project information, department, working hours, after-hours activities, system interactions, and user roles are just a few of the many organisational activities captured by the 693,649 records and about 830 behavioural features that make up the dataset. These characteristics of behaviour provide enough information to differentiate between typical employee actions and questionable insider activity. In order to ensure that the dataset is free of outliers, missing values, and inconsistencies before developing the model, this ensures that the data used for machine learning analysis is reliable.

The CERT dataset has a significant class imbalance, which is one of the main obstacles to insider threat identification. There are 693,649 documents in all, but only 1,307 of them pertain to insider threats; the rest are just regular business. Because of this bias, traditional machine learning algorithms start to favour the majority, making them less effective at detecting insider assaults, which are very uncommon but crucial nevertheless. Using the Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic minority-class samples to equalise the training dataset, the suggested approach gets around this issue. By including SMOTE, machine learning models may acquire representative traits from both benign and malevolent actions, which greatly enhances their ability to identify insider threats.

If the data is balanced, the next step is to split the dataset in half. Half will be used for training, while the other half will be used for testing and model assessment. To facilitate fair comparison of performance, many ML algorithms are trained using the same preprocessing parameters. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, AdaBoost, and XGBoost are some of the classifiers that have been studied. When it comes to behavioural categorisation, every algorithm has its own set of benefits. Logistic Regression offers a model with a probability foundation, Decision Trees conduct hierarchical decision analysis, Random Forest enhances prediction robustness with ensemble learning, Support Vector Machines (SVM) find intricate decision boundaries, KNN uses neighbourhood similarity to categorise user behaviour, Naïve Bayes applies probability to classification, AdaBoost helps weak learners with adaptive boosting, and XGBoost optimises gradient boosting for high-dimensional behavioural data.

Every classifier is hyperparameter optimised before training in order to have the best possible prediction performance. An suitable maximum tree depth and learning rate are specified with XGBoost, whereas Decision Trees use the Gini impurity criteria, Random Forest uses 100 decision trees, KNN uses k = 5, and AdaBoost uses 50 estimators. Ensemble learning techniques, which examine very complicated patterns of organisational behaviour, benefit greatly from hyperparameter adjustment, which increases model generalisability while decreasing overfitting.

Lastly, the efficacy of each trained model in identifying insider threats is assessed using Accuracy, Precision, Recall, and F1-score. The best classifier for cybersecurity applications in a business may be objectively determined via comparative study. When it comes to intelligent insider threat detection using ensemble machine learning techniques—while maintaining the original methodology, dataset, algorithms, and experimental design—the experimental evaluation shows that Random Forest and AdaBoost achieve the highest classification performance with a 97.5% accuracy.

IV. CERT DATASET PREPARATION AND MODEL DEVELOPMENT

An insider threat detection system's efficacy is highly dependent on the dataset's quality and the classification machine learning models' resilience. For the purpose of detecting harmful insider activity, the CERT 5.2 Insider Threat Dataset is the principal source of behavioural information in the suggested architecture. User logins, file transfers, USB device usage, project assignments, department, working hours, after-hours activities, system interactions, and various organisational activities are represented in the dataset's 693,649 records with about 830 behavioural features. These detailed behavioural characteristics allow machine learning algorithms to accurately differentiate between normal and questionable employee actions.

The acquired dataset is subjected to thorough preprocessing prior to model training in order to guarantee dependable learning and enhance data quality. The dataset is first checked for inconsistent observations, duplicate entries, and missing values. According to the results, there are no major outliers or missing values in the CERT dataset that would need a lot of manual adjustment. Machine learning algorithms are able to analyse behavioural information effectively when categorical qualities are encoded into numerical representations. By standardising the dataset, these preprocessing techniques enhance the stability of the model during assessment and training.

The very skewed distribution of classes in the CERT dataset is one of the main problems with it. The remaining 693,649 observations represent typical employee behaviour, while only 1,307 records correspond to insider threats. Because the models start to lean toward predicting typical activities due to this imbalance, traditional machine learning algorithms become much less effective at identifying instances involving minority classes. Using the Synthetic Minority Oversampling Technique (SMOTE), the suggested framework is able to overcome this obstacle. By balancing the training dataset and improving machine

learning models' ability to detect rare but critically important malicious activities, SMOTE generates synthetic minority-class samples by interpolating existing insider threat instances.

The produced dataset is then split in half: 80% for training and 20% for testing, after data balancing. Machine learning models are built using the training dataset, and their ability to classify unseen behavioural patterns is evaluated using the testing dataset. This partitioning strategy reduces the likelihood of overfitting and provides an unbiased evaluation of the generalisability of the model. To ensure a fair comparison of model performance throughout the experimental analysis, it is important to maintain identical training and testing conditions for every classifier.

To find the best method for detecting insider threats, the suggested framework examines eight different machine learning algorithms. A probabilistic classifier for binary classification can be established as a baseline using logistic regression. Decision Tree constructs hierarchical decision structures using behavioral features, while Random Forest combines multiple decision trees to improve prediction robustness and reduce overfitting. The ideal decision boundaries for distinguishing between malicious and legitimate activities are identified by Support Vector Machine (SVM). K-Nearest Neighbors (KNN) classifies user behavior by examining neighboring activity patterns, whereas Naïve Bayes performs probabilistic classification under feature independence assumptions. And XGBoost uses gradient boosting to maximise classification accuracy for high-dimensional organisational data, while AdaBoost uses adaptive boosting to bolster weak learners. Each algorithm contributes unique strengths toward identifying insider threats under different behavioral conditions.
To maximize classification performance, hyperparameter optimization is performed for every machine learning model. Random Forest is configured using 100 decision trees, Decision Tree applies the Gini impurity criterion, KNN operates with k = 5, AdaBoost utilizes 50 estimators with an optimized learning rate, and XGBoost is configured using an appropriate maximum depth and learning rate. These optimized configurations improve prediction accuracy, enhance generalization capability, and reduce overfitting when analyzing complex behavioral data generated within organizational environments.

V. PERFORMANCE EVALUATION

A thorough comparison of eight machine learning algorithms utilising the CERT 5.2 Insider Threat Dataset is conducted to assess the performance of the proposed framework for detecting insider threats. The pretreatment conditions used to train each classifier are the same. These include optimising the hyperparameter values, an 80:20 split between training and testing, and data balancing using SMOTE. By using Accuracy, Precision, Recall, and F1-score, we can objectively compare the effectiveness of each model in detecting harmful insider behaviours in organisational settings.

Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, AdaBoost, and XGBoost are the main areas of attention in the first assessment. Each classifier looks for characteristics of user behaviour that can indicate an insider threat, such as logins, file access operations, USB use, project assignments, departmental information, and actions performed outside of normal business hours. Due to differences in learning processes and decision limits, all models effectively categorise organisational activities; nevertheless, their detection capacities range significantly. Due to its capacity to decrease overfitting and increase prediction stability via the combination of several decision processes, ensemble learning algorithms often exhibit better performance.

The experimental findings show that the classifiers that were assessed varied significantly from one another. Out of all the models we looked at, Logistic Regression had the worst overall performance with an accuracy of 58%. Classification capacity is enhanced by Naïve Bayes at 75% accuracy, Support Vector Machine (SVM) at 86% accuracy, and K-Nearest Neighbours (KNN) at 90% accuracy. Decision Tree and XGBoost, both of which are based on trees, achieve 97% accuracy, which is a significant improvement. Random Forest and AdaBoost yield the best results, with an accuracy of 97.5% each. This shows that ensemble learning approaches are useful for identifying insider threats in datasets that are very skewed inside organisations.

The suggested approach uses classification accuracy, Precision, Recall, and F1-score to assess each model. By comparing each classifier's accuracy in detecting insider threats while limiting false positives, these metrics provide a thorough picture of prediction quality. Decision Tree, Random Forest, AdaBoost, and XGBoost outperform traditional classification algorithms in terms of F1-scores, as shown by the experimental study, which shows that these algorithms routinely attain high recall and accuracy values. These results show that ensemble learning approaches are better suited to real-world insider threat detection applications due to their balanced classification performance, which is crucial for both accuracy and dependability.

VI. PRACTICAL DEPLOYMENT

The suggested methodology for detecting insider threats uses machine learning to automatically spot unusual patterns of behaviour that might be an indication of an insider attack, allowing for continuous surveillance of organisational resources and employee actions. The suggested architecture constantly examines user behaviour using trained machine learning models to detect any dangers in real time, in contrast to conventional security systems that depend mostly on preset rules or signature-based detection. Organisations may improve their cybersecurity with this smart monitoring method, which

lowers the risk of data leakage, harmful insider actions, and unauthorised access.

Login records, file access operations, interactions with USB devices, project assignments, departmental information, working hours, and after-hours activities are just a few examples of the user activities that are continuously collected and supplied to the insider threat detection framework in a practical deployment environment. Data validation, categorical encoding, feature preparation, and behavioural normalisation are some of the preprocessing processes that are performed on the incoming data before prediction. These activities are also utilised during model building. Prediction reliability is enhanced by maintaining consistency between the training environment and real-time operating settings by the use of same preprocessing processes during deployment.

To address the significant class imbalance in the CERT dataset, the framework uses the Synthetic Minority Oversampling Technique (SMOTE) during model construction. This is done since insider threat occurrences are inherently infrequent. Without additional balance needed during prediction, the trained machine learning models use the learnt behavioural patterns to categorise newly observed organisational activities. Under real-world organisational situations, this method allows the system to accurately discern between typical employee behaviour and possibly malevolent insider behaviours, all while maintaining high classification accuracy.

With the use of many trained machine learning models, such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, AdaBoost, and XGBoost, the operational detection process assesses user actions all at once. Classifiers evaluate user actions based on a variety of behavioural variables to determine whether they are indicative of genuine organisational behaviour or potential insider threats. Experimental results show that Random Forest and AdaBoost have better detection capabilities than other methods. This makes them ideal for use in business security monitoring systems that need reliable and accurate detection.

Security teams may analyse unusual actions before they become major cybersecurity problems since the framework instantly creates security notifications for administrators whenever suspicious behaviour is observed. In order to react quickly and minimise possible financial losses or data breaches, organisations may identify strange login patterns, excessive file transfers, unauthorised USB use, or aberrant after-hours activity. Forensic investigation and long-term security planning are both aided by the produced prediction reports, which provide comprehensive details about behavioural abnormalities and model forecasts.

Therefore, the suggested methodology offers a workable and extensible way to improve cybersecurity in organisations by detecting intelligent insider threats. The framework improves security monitoring while keeping the original methodology, dataset, methods, and experimental findings from the source paper. It does this by combining behavioural analysis, SMOTE-based model creation, numerous machine learning classifiers, and real-time prediction. Its implementation may help businesses with proactive cybersecurity management, insider risk reduction, and threat visibility.

VII. CONCLUSION

Because harmful operations are frequently carried out by trusted personnel with authorised access to organisational resources, insider threats continue to be one of the most critical cybersecurity issues. When it comes to spotting insider assaults, traditional security measures aren't great since they mostly use signature-based detection approaches and established criteria. This research devised a framework for intelligent behavioural analysis and automated threat identification using the CERT 5.2 Insider Threat Dataset, which is based on machine learning, to overcome these constraints. Maintaining the study's original approach, the suggested framework enhances insider threat detection via data pretreatment, SMOTE-based class balance, and several machine learning methods.

Eight machine learning classifiers, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, AdaBoost, and XGBoost, were compared in the experimental inquiry. All of these classifiers were evaluated using the same preprocessing settings. To enable objective comparison across various categorisation strategies, model performance was tested using Accuracy, Precision, Recall, and F1-score. Random Forest and AdaBoost outperformed the other machine learning algorithms in terms of classification accuracy (97.5%), recall (80%), and F1-score (85%), according to the experimental data. The results show that ensemble learning methods work well to detect insider threats' complicated behavioural patterns.

To improve prediction performance, the CERT dataset had a severe class imbalance, but the Synthetic Minority Oversampling Technique (SMOTE) helped fix it. The SMOTE algorithm successfully trained machine learning models to distinguish between legitimate business operations and those perpetrated by malevolent insiders by creating synthetic minority-class examples. Organisational cybersecurity and decision-making were both improved by the suggested framework's enhanced capacity to identify infrequent insider threat occurrences with little majority-class classification bias.

All things considered, the suggested ML framework offers a dependable, scalable, and effective method for detecting insider threats to organisations via smart behavioural analysis and ML comparison. If implemented properly, the framework may help security managers spot questionable user behaviour, lessen the likelihood of cyberattacks, and fortify their company's security monitoring infrastructure. While this study presents a solid foundation for machine

learning, future research could look into deep learning models, explainable AI techniques, advanced feature selection methods, and hybrid ensemble learning strategies to enhance the accuracy and adaptability of insider threat detection in large-scale organisational settings.

REFERENCES

1. S. G., S. Silas, and E. B. Rajsingh, "A Pragmatic Enquiry to Learn Recent Trends in Insider Threat Detection Approaches," in Proc. 7th Int. Conf. Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2024.
2. F. Femi-Oyewole et al., "Survey on Predictive Algorithms to Detect Insider Threat on a Network Using Different Combination of Machine Learning Algorithms," in Proc. Int. Conf. SEB4SDG, Omu-Aran, Nigeria, 2024.
3. H. E.-E. Abdallah et al., "Performance Evaluation Framework for Insider Threat Detection Using Machine Learning," in Proc. Intelligent Methods, Systems and Applications (IMSA), Giza, Egypt, 2024.
4. D. Sridevi et al., "Detecting Insider Threats in Cybersecurity Using Machine Learning and Deep Learning Techniques," in Proc. Int. Conf. Communication, Security and Artificial Intelligence, Greater Noida, India, 2023.
5. R. Yousef et al., "A Machine Learning Framework and Development for Insider Cyber-Crime Threat Detection," in Proc. Int. Conf. Smart Applications, Communications and Networking, Istanbul, Türkiye, 2023.
6. N. Dixit et al., "Insider Threat Classification Using KNN Machine Learning Technique," in Proc. Int. Conf. Contemporary Computing and Communications, Bangalore, India, 2023.
7. Y. Jing et al., "Analyze Organizational Internal Threats Using Machine Learning," in Proc. Int. Conf. Applied Machine Learning, Dalian, China, 2023.
8. B. Nagabhushana Babu and M. Gunasekaran, "An Analysis of Insider Attack Detection Using Machine Learning Algorithms," in Proc. IEEE 2nd Int. Conf. Mobile Networks and Wireless Communications (ICMNWC), Tumakuru, India, 2022.
9. S. K. Mamidanna et al., "Detecting an Insider Threat and Analysis of XGBoost Using Hyperparameter Tuning," in Proc. Int. Conf. Advances in Computing, Communication and Applied Informatics, Chennai, India, 2022.
10. X. Sun et al., "Insider Threat Detection Using an Unsupervised Learning Method: COPOD," in Proc. Int. Conf. Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021.
11. R. Nasir et al., "Behavioral Based Insider Threat Detection Using Deep Learning," IEEE Access, vol. 9, pp. 143266–143274, 2021.
12. P. Varsha Suresh et al., "Insider Attack: Internal Cyber Attack Detection Using Machine Learning," in Proc. 12th Int. Conf. Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021.
13. D. C. Le et al., "Analyzing Data Granularity Levels for Insider Threat Detection Using Machine Learning," IEEE Transactions on Network and Service Management, vol. 17, no. 1, pp. 30–44, Mar. 2020.
14. C. C. Aggarwal, Data Mining: The Textbook. Cham, Switzerland: Springer, 2015.
15. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
16. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
17. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
18. C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
19. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
20. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
21. P. Indyk and R. Motwani, "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality," in Proc. 30th ACM Symp. Theory of Computing (STOC), 1998, pp. 604–613.
22. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
23. P. Ramesh Babu, "Automated Student Document Analysis and Query System Using OCR and Conversational AI," pp. 358–366, 2026.
24. P. Ramesh Babu, "Machine Learning-Based Insider Threat Detection Using CERT Dataset," 2026.