Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCED INSIDER THREAT DETECTION THROUGH MACHINE LEARNING APPROACH WITH IMBALANCED DATA RESOLUTION

PENNADA SIVA SATYA PRASAD[1], SASMITA KUMARI NAYAK[2], Dr. M. VAMSI KRISHNA[3]

[1]Research Scholar, Department of CSE, Centurion University of Technology and Management, Bhubaneswar, Odisha and Assistant Professor, Aditya Engineering College, Surampalem, India

[2]Associate Professor, Department of CSE, Centurion University of Technology and Management, Bhubaneswar, Odisha, India.

[3]Professor, Department of IT, Aditya Engineering College, Surampalem, Andhra Pradesh, 533437, India

E-mail: [1]sivasatyaprasadp@gmail.com, [2]nayaksasmita484@gmail.com, [3]vkmangalampalli@gmail.com

ABSTRACT

An insider threat is the risk that person inside an organization may pose to the company's security, data, or resources. Insider threat detection is a crucial component of a comprehensive cybersecurity strategy. By identifying and mitigating risks from within the organization, businesses can better protect their assets, maintain trust, and ensure compliance with legal and regulatory requirements. This paper addresses the detection of insider threats using machine learning algorithms. A famous CERT dataset was used for the experiments. The collected dataset is largely imbalanced. The ML algorithms cannot perform well with imbalanced datasets. So, data imbalance can be resolved by using three over sampling techniques namely random oversample, smote, adasyn and three under sampling techniques namely random under sample, Cluster centroids and Edited Nearest Neighbors. Later, five ML algorithms namely Logistic Regression, Adaboost, Decision Tree, Random Forest and Naïve Bayes applied to the datasets generated through over sampling and under sampling techniques. To further increase the performance of the model, an ensemble learning is proposed along with principal component analysis. The experimental results demonstrated that the proposed model surpassed the performance of existing models for insider threat detection.

Keywords: *Insider Threat Detection, Data Imbalance, Over Sampling, Under Sampling, Machine Learning, Ensemble Learning.*

1. INTRODUCTION

In the current cybersecurity landscape, the importance of safeguarding sensitive information and digital assets against diverse threats is more pronounced than ever. While external threats such as hackers and malware remain a concern, insider threats, those originating from individuals within an organization, have emerged as a formidable challenge. These threats encompass a range of activities, from data theft and unauthorized access to sabotage and espionage, and they can have severe consequences for organizations, including financial losses, damage to reputation, and regulatory penalties. Insider threats are particularly insidious because they can exploit their legitimate access to systems and data, making them harder to detect using traditional security measures. To address this growing concern, research in the domain of detecting and mitigating insider threats has gained significant prominence. The increasing incidence of insider threats has become a pressing concern for organizations, prompting the exploration of advanced solutions to effectively detect and prevent such threats. The complexity of insider threat scenarios adds to the challenge, as these threats manifest in diverse forms, making it difficult to identify them through traditional rule-based or signature-based systems. Machine learning emerges as a promising approach to address this challenge by recognizing anomalous patterns and behaviors associated with insider threats. Insider

threats pose a significant and growing challenge to the cybersecurity of organizations, as malicious or unintentional actions by individuals with authorized access can lead to data breaks, monetary losses, and mutilation to an organization's status. Traditional methods of detecting insider threats often fall short due to their reliance on static rules and signatures, which are ineffective at identifying evolving and subtle insider threat behaviors. As a result, there is a pressing need to develop and evaluate more sophisticated and proactive approaches for the classification and detection of insider threats using ML algorithms. Recent advancements in machine learning, including techniques like DL and ensemble methods, provide organizations with opportunities to enhance the accuracy and efficiency of insider threat detection. Recognizing the need for proactive protection, organizations are moving beyond reactive measures to implement strategies that identify and alleviate insider threats earlier they can cause harm. The landscape of compliance and regulatory requirements further accentuates the importance of robust insider threat detection capabilities. Many industries and organizations face stringent mandates, and non-compliance can lead to severe consequences. Protecting intellectual property, proprietary information, and sensitive data has become crucial for organizations seeking to maintain a competitive edge and safeguard their interests. The public awareness of high-profile insider threat incidents has grown, drawing attention to the potential harm they can cause. This heightened awareness places increased pressure on organizations to enhance their security measures and invest in effective insider threat detection mechanisms.

Machine learning offers a promising approach to recognize anomalous patterns and behaviors associated with insider threats. Recent advancements in ML, including DL and ensemble methods, present an opportunity to augment the accuracy and efficacy of insider threat detection. Organizations are increasingly recognizing the need to go beyond reactive measures and implement proactive strategies for identifying and mitigating insider threats before they cause harm. Many industries and organizations are subject to stringent compliance and regulatory mandates that necessitate robust insider threat detection capabilities. Failure to comply with these requirements can result in severe consequences. Intellectual property, proprietary information, and sensitive data are valuable assets for many organizations. Safeguarding these assets against insider threats is imperative for maintaining a competitive edge and protecting the organization's interests. High-profile insider threat incidents have garnered public attention, raising awareness about the potential harm they can cause. This increased awareness places pressure on organizations to enhance their security measures.

2. LITERATURE SURVEY

Machine Learning, a subset of AI, enables computers to perform tasks without explicit programming through the use of algorithms and statistical models [1]. It is not a new practice to use ML and DL methods in the context of cyber security applications. In [2], authors used deep feature synthesis to create 70,000 user-specific features based on historical data to identify behavioral tendencies. PCA reduced dimensionality and enhanced machine learning, discovering insider risks. Categorization and anomaly detection methods were included. The anomaly detection model was 91% accurate. The CERT insider threats dataset was used to demonstrate SMOTE balancing's ability to reduce dataset imbalance. Recall and accuracy rose, while precision fell. The SVM model and feature extraction procedure surpassed all other machine learning models with 100% classification accuracy. The authors in [3] identified research potential for insider threat identification using machine learning algorithms. They conducted a systematic literature review that required careful planning, selecting, extracting, and analyzing the data. In order to enhance insider threat solutions, the three detection techniques combination, selection, and execution were examined and recommendations were made for further study. The authors in [4] introduced DL hybrid LSTM models that used Google's Word2vec, LSTM, and GLoVe to fill insider threat detection gaps. These hybrid DL models were tested against cutting-edge ML models including XGBoost, AdaBoost, RF, KNN, and Logistics Regression. The inquiry sought to solve genuine dataset, accuracy, and false alarm issues. Unexpectedly, ML-based models detected insider threats better than DL-based models. Previous investigations revealed the proposed technique was efficient utilizing a real dataset.

Current research advancements were applied to discover insider risks in [5]. Google's Word2vec LSTM GLoVe LSTM was coupled with the first two DL hybrid LSTM models. Second, two hybrid DL models were compared to top ML models as XGBoost, AdaBoost, RF, KNN, and LR. Thirdly, our research employed an actual dataset,

was accurate, and greatly minimized false alarms. ML models outperformed DL. Based on earlier research, the results identified insider dangers using real data. Insider threat detection technique based on self-supervised and ensemble learning was presented by the authors in [6]. To improve the efficacy of detection, they also devised an entity representation technique based on TF-IDF. The suggested approach was able to successfully identify malicious sessions in the CERT4.2 and CERT6.2 datasets, according to experimental findings. The author of [7] suggested an optimization-based insider threat detection approach. Spider Monkey Optimization was used to identify attitudes in Carnegie Mellon University's CERT R4.2 dataset for insider threat detection. The dataset was downloaded and extracted, then pre-processed to remove noise and reject null values. Content field and Natural Language Processing toolkit extracted features. Spider Monkey Optimization was used to choose features using Linear Discriminant Analysis' contribution factor. The TextBlob library calculated the polarity of the greatest contribution LDA document picked by SMO.

A double-layer architecture for Medical Image Tampering (MIT) detection was presented in [8]. The first layer integrated, transformed, and sampled data. From eight sample methods, Nearmiss2 (NM-2) performed best and was chosen. The second layer used NM-2 sampled data in an abnormal MIT detection model using several anomaly detection methods. In anomaly detection, the approach addressed the Counterfeit Image Problem (CIP). MIT detection was acceptable with the suggested double-layer architecture, which used NM-2 and One-class SVM, achieving 79% f-score and 83% accuracy. The authors presented a mouse biobehavioral and deep learning-based user authentication approach in [9] to correctly and efficiently authenticate existing computer users to counter insider risks. The strategy worked in experiments with 10 users using an open-source dataset. The approach performed a user authentication job every 7 seconds with 2.94% false acceptance and 2.28% false rejection. The survey in [10] reviewed frequently used insider threat detection datasets and new deep learning research. The results showed that deep learning models outperformed typical ML techniques in insider threat identification. Deep learning faced problems including limited labeled data and adaptive attacks when used to advance this goal. The survey addressed these issues and suggested further research to improve deep learning for insider threat identification.

The authors created an unsupervised ML algorithm to detect dangerous insider behaviors utilizing data from many technical sources in [11]. The system, which was easy to build, was tested using existing machine learning techniques and recognized dangerous insider activity during training but not during testing. These results suggested that machine learning may help identify insider dangers, but not entirely. System performance was improved by include file names, email subjects and headers, and web site types. The authors in [12] suggested a methodology for insider threat identification by semi-supervised and supervised ML, data stream examination. The algorithms were Isolation Forest, Elliptic Envelope, and Local Outlier Factor. Results were assessed and achieved good precision, recall, and F1-measure values. The authors in [13] recommended a novel multiple security log approach. Texts from security logs formed a corpus. The corpus-trained Word2vec approximated insider behavior posterior probability. Converted events with behavioral likelihood below a threshold were suspicious, and a user was malicious if they had several. The testing using CERT Programs internal threat database v6.2 revealed that the recommended method was effective and scalable and advised parameter and threshold changes. DDoS assaults are identified using a multiclass dataset that includes Smurf, SIDDoS, HTTP-Flood, and UDP-Flood [14]. To maintain impartiality, training and testing use balanced datasets. Four experiments with varying reduced characteristics are run. The authors in [15] applied KNN model for insider threat detection and acquired good detection rate. For insider threat detection, the KNN Classification Algorithm [16] categorized people as genuine or not genuine. A threat detection approach using face recognition and surveillance was also shown. In [17], the authors tested a ML-based user-centered insider threat detection system. Machine learning detected harming acts and insiders by analyzing data at various granularities in actual circumstances. Multiple performance indicators were utilized to measure system performance after a thorough examination of typical insider threat scenarios. Test results demonstrated that the machine learning-based detection method could identify new hostile insiders in unseen data given limited ground truth.

In [18], system log characteristics were condensed to retain essential information and provide correct insight. Two unsupervised algorithms were evaluated for insider threat

identification using daily and frequently aggregated system log formats. Innovatively, the anomaly score from the previous cycle was employed as each user's trust score in the next cycle's model, suggesting its efficacy in detecting insiders. The model initially detected insider threats using user psychometric scores. The proposed technique beat previous methods on the CERT insider threat dataset. The authors in [19] classified insider threats using supervised, unsupervised, and reinforced machine learning. A technical data-based unsupervised ML system identified dangerous insider behaviors. Some detrimental insider behavior was recognized during training but not during testing. It was shown that ML can detect insider dangers. A methodology for identifying damaging insider threats using SVM was provided [20]. The highest anomaly ratings were utilized to categorize, forecast, and identify harmful and non-malicious actions. Experimental findings showed that the suggested system could identify malicious insiders with much higher anomaly scores than regular users in [21], user behavior profiling was used to identify insider threats. The ensemble hybrid machine learning technique used MSLSTM and CNN to identify time series anomalies. Multistate LSTM beat single-state LSTM in the investigation. When trained with a publicly accessible insider threat dataset, Multistate LSTM detected insider threats with an AUC of 0.9 on the train data and 0.9 on the test data.

Most of the previous works applied conventional ML and DL algorithms for insider threat detection. The problem with conventional algorithms is that the conventional models cannot able to analyze the features properly due to algorithm restrictions. One more important issue identified from the previous works is that, most of the works directly used the datasets. But, the datasets for insider threat detection may contain imbalanced class labels. The performance achieved with this imbalanced dataset is reliable. In this work, both these issues are resolved.

In this paper, the authors proposed ML based model for insider threat detection. Before applying ML model, dataset imbalance issue resolved through oversampling techniques SMOTE and ADASYN. The rest of the paper is organized as follows: Section 2 describes the proposed approach. Results and discussion are presented in Section 3. Conclusion is presented in Section 4.

.

3. RESEARCH METHODOLOGY

The proposed method for insider threat detection is shown in Figure 1.

3.1. Collection of Dataset and Preprocessing

Initially, CERT insider threat dataset was collected [22]. The dataset contains 6,93649 with 830 features. The target variable in the dataset is “insider”. The collected dataset is verified for missing values and outliers and the dataset is not having those issues.

3.2. Handling Class Imbalance

Out of 6,93649 samples of data, the number of samples with insider threat class label 0 are 6,92342 and number of samples with class label insider threat as 1 are 1,307. So, this is imbalanced dataset. If this dataset directly used for classification problems, then the results are irrelevant and inaccurate. There are two strategies to solve this problem. One strategy for dealing with datasets that lack balance is to oversample the classes that are underrepresented. To resolve class imbalance, three over sampling techniques namely “Random Oversampling”, “SMOTE” and “ADASYN” are used. Another strategy for solving class imbalance is under sampling. Two under sampling techniques namely Cluster Centroids, Edited Nearest Neighbors.

3.2.1 Random oversampling

Random oversampling is a simple way to correct class Out of 6,93649 samples of data, the number of samples with insider threat class label 0 are 6,92342 and number of samples with class label insider threat as 1 are 1,307. So, this is imbalanced dataset. If this dataset directly used for classification problems, then the results are irrelevant and inaccurate. There are two strategies to solve this problem. One strategy for dealing with datasets that lack balance is to oversample the classes that are underrepresented. To resolve class imbalance, three over sampling techniques namely “Random Oversampling”, “SMOTE” and “ADASYN” are used. Another strategy for solving class imbalance is under sampling. Two under sampling techniques namely Cluster Centroids, Edited Nearest Neighbors imbalance by boosting minority class occurrences. To balance class distribution, minority class instances are randomly selected and duplicated. It duplicates minority class occurrences in the training dataset to boost their representation.

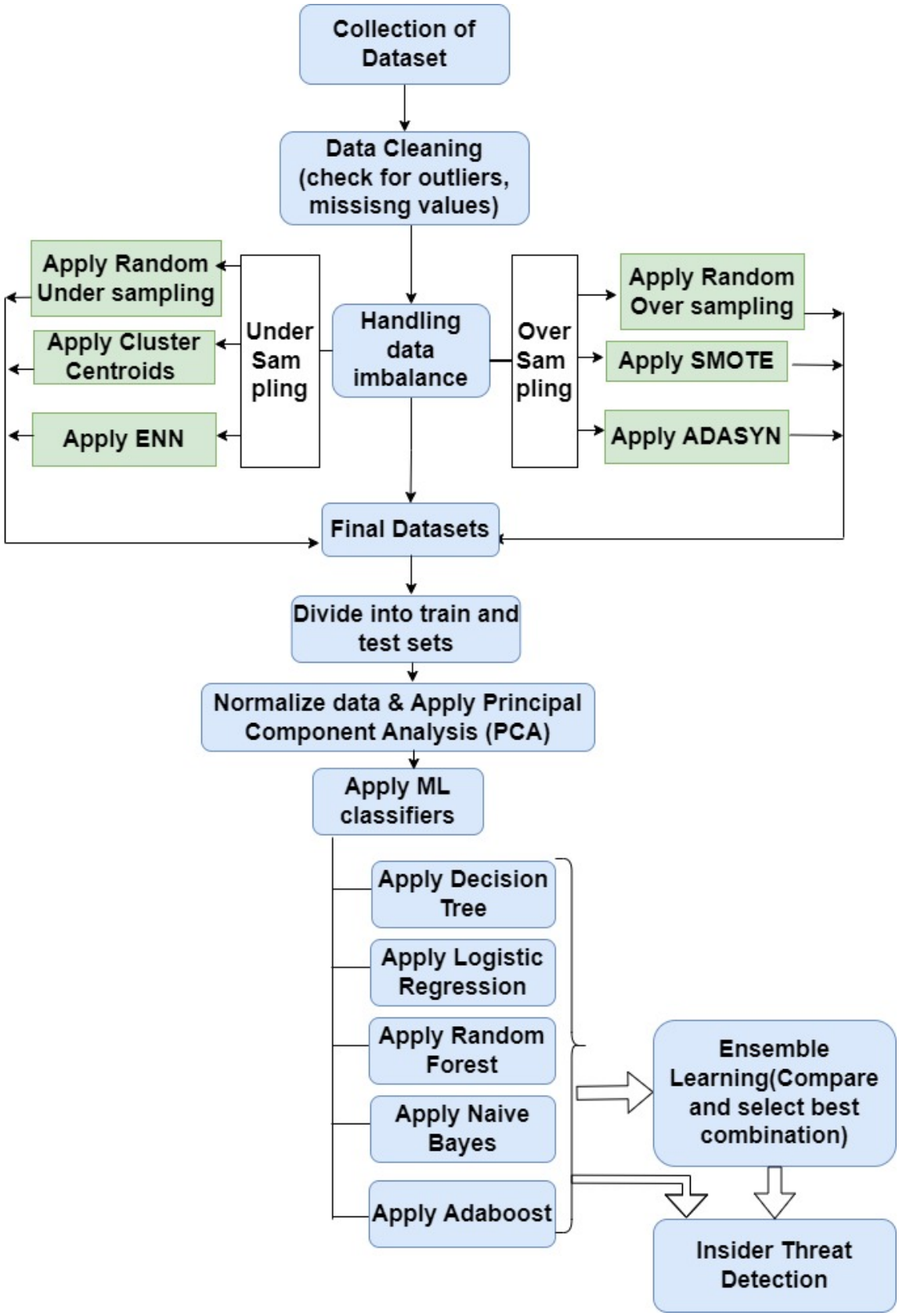

Figure 1: Proposed Method for insider threat detection

3.2.2 SMOTE

SMOTE (Synthetic Minority Oversampling Technique) is used in the field of ML to resolve the issue of class imbalance. This method is employed most often in classification tasks where one class considerably outnumbers the other. It is possible for class imbalance to cause machine learning models to have a bias toward the class that is in the majority, which may hinder their capacity to properly forecast the class that is in the minority. The primary objective of SMOTE is to generate synthetic samples for the minority class in order to achieve a more equitable distribution of classes. The process involved for sampling is as follows: Initially, a minority class instance is chosen. Subsequently, the algorithm identifies the k-nearest neighbors to this selected instance (k=3 in this paper). From the identified neighbors, one is then randomly selected. The final step involves creating a synthetic instance by blending the features of the initially chosen instance with those of the randomly selected neighbor. This methodology effectively introduces synthetic instances to the minority class, contributing to a more balanced distribution in the dataset. After applying SMOTE the number class labels for insider thread 0 and 1 are changed to 6,92,342. Now the dataset became balanced.

3.2.3 ADASYN (Adaptive Synthetic Sampling)

setting of unbalanced datasets. ADASYN stands for Adaptive Synthetic Sampling. ADASYN tackles the problem of uneven class distribution by producing synthetic examples for the underrepresented classes in a way that is both more data-driven and flexible. ADASYN addresses unbalanced machine learning datasets data-drivenly. ADASYN dynamically analyzes instance density in feature space, unlike traditional oversampling. This flexibility lets the system generate synthetic samples in places with smaller minority class density, adjusting the oversampling process to the dataset's local features. An important feature of ADASYN is significance weighting. This includes weighting minority class instances by learning difficulty. Hard-to-classify instances are weighted higher. The oversampling technique prioritizes synthetic samples for cases that improve the model's minority class performance based on this strategic weighting. It generates synthetic samples like SMOTE. Interpolating minority class instances creates synthetic instances. The algorithm's concentration on low-density zones and instance difficulty make synthesis more flexible and sophisticated.

3.2.4. Random Under sampling

The purpose of this method is to balance the class distribution by reducing the number of samples in the majority class (the class with more instances) randomly until the class distribution is more balanced. This technique can be useful when dealing with classification tasks to avert the system from being unfair toward the majority class.

3.2.5. Cluster Centroids Under sampling

The Cluster Centroids method aims to balance the class distribution by replacing the majority class clusters with their centroids. This is accomplished in a manner that reduces the number of majority class samples in the new dataset while preserving the overall distribution of the majority class.

.3.2.6. Edited Nearest Neighbors (ENN)Under sampling

ENN method aims to clean the dataset by removing instances of the majority class whose class label differs from the majority class of their k-nearest neighbors. It focuses on removing potentially noisy samples that may have been mislabeled.

3.3. Applying ML models

After applying over sampling & under sampling techniques, class imbalance issue is resolved. As three over sampling techniques and three under sampling techniques applied, now six different datasets available. But random oversampling and random under sampling datasets are not considered for analysis. So, four datasets available. Next, several ML algorithms were applied to these four datasets. Before applying ML models, the final datasets are normalized. As there are more than 800 features in the dataset, the complexity of model is high. To handle this, we applied PCA before applying ML techniques. The ML techniques used this work are Logistic Regression, Decision Tree, Random Forest, Adaboost and Naïve Bayes. After applying ML techniques results are noted and later ensemble techniques applied for further increase in performance of the model.

3.3.1 Decision tree

The ML approach of decision tree classification utilizes a tree-like structure to make judgments or predictions. The process involves recursively partitioning the dataset based on features to maximize class separation and construct the tree. At each node, a feature determines a decision, leading to branches representing different outcomes. This recursive process continues until a stopping condition, such as reaching a specified

tree depth or a minimum leaf node sample count. Decision trees are known for their interpretability and effectiveness in labelling cases based on the values of their features.

3.3.2 Random Forest

It is an ensemble learning approach employed for categorization, utilizing bootstrapping sampling to generate multiple decision trees with unpredictable node properties. The final prediction is determined through a regression average or a majority vote from the classification trees. The inherent unpredictability and diversity in Random Forests contribute to their effectiveness in machine learning, reducing overfitting and enhancing model robustness. Moreover, the algorithm provides insights into model behavior by highlighting the significance of features in the analysis.

3.3.3 Logistic regression

Logistic regression classification model's binary outcome probability in ML. Logistic regression predicts class probability, unlike linear regression, which predicts continuous values. It uses the logistic function (sigmoid function) to convert a linear combination of input information into a positive class probability between 0 and 1. The final categorization is based on a threshold. Due to its simplicity, interpretability, and effectiveness, logistic regression is preferred for binary classification.

3.3.4 Naïve bayes

It is a probabilistic machine learning method that categorizes instances into predefined classes. Built on Bayes' theorem, it assumes feature independence given the class label, earning the "naive" label.

3.3.5 Ensemble Learning

Ensemble learning uses several model predictions to produce a better, more robust model. Ensemble methods can improve generalization, reduce overfitting, and enhance predictive performance. Two common ensemble techniques are stacking and voting classifiers. In this paper, both stacking and voting classifiers applied.

4. EXPERIMENTS AND RESULTS

4.1 Applying ML algorithms

The original dataset comprises 6,93,649 samples with 6,92,342 samples with label 0 and 1307 samples with label 1. In over sampling, the number of samples of 0 are increased. In under sampling, the number of samples of label 0 is decreased. In ENN, the number of total samples changes based on the dataset. For all the four datasets, training and testing set ration is 80:20. The number of samples in each dataset and number of train and test samples details are shown in Table 1. The proposed algorithms applied with four datasets created from SMOTE (Dataset-1) and ADASYN(Dataset-2), Cluster Centroids (Dataset-3) and ENN(Dataset-4).

Table 1: Dataset division

Data set	Total samples	Training Samples	Testing Samples
Data set-1	13,84,684	11,07,748	2,76,936
Data set-2	13,84,684	11,07,748	2,76,936
Data set-3	2,614	2091	523
Data set-4	58,409	46,727	11,682

4.1.1. Apply ML algorithms with dataset-1

The proposed five algorithms are applied for Dataset-1(created through SMOTE technique). The results after applying algorithms are shown in Table I.

From Table-2, it is identified that RF, adaboost and decision tree given good results with dataset-1 for insider threat detection. The precision, recall and accuracy all are good for these three algorithms.

4.1.2. Apply ML algorithms with dataset-2

The proposed five algorithms are applied for Dataset-2(created through ADASYN technique). The results after applying algorithms are shown in Table 3.From Table-3 and Figure 3, it is evident that RF, adaboost and decision tree given good results with dataset-2 for insider threat detection. The precision, recall and accu3.racy all are good for these three algorithms.

4.1.3. Apply ML algorithms with dataset-3

The proposed five algorithms are applied for Dataset-3(created through Cluster Centroids technique). The results after applying algorithms is shown in Table 4. Table-4 reveals that random forest, Adaboost, and decision tree achieved favorable outcomes in detecting insider threats with dataset-3. These three algorithms demonstrated satisfactory precision, recall, and accuracy.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Results with five ML models for Dataset-1

Algorithm	Class	P	R	F1	Accuracy
Naïve Bayes	NO	68%	91%	78%	74%
	YES	86%	58%	69%	
Logistic Reg	NO	57%	78%	66%	59.4%
	YES	65%	41%	50%	
Decision Tree	NO	98%	97.9%	98%	98%
	YES	97.9%	98%	98%	
Random Forest	NO	98%	98.1%	98%	98%
	YES	98%	98.2%	99%	
Adaboost	NO	97.6%	98%	98%	98%
	YES	98.3%	98%	98%	

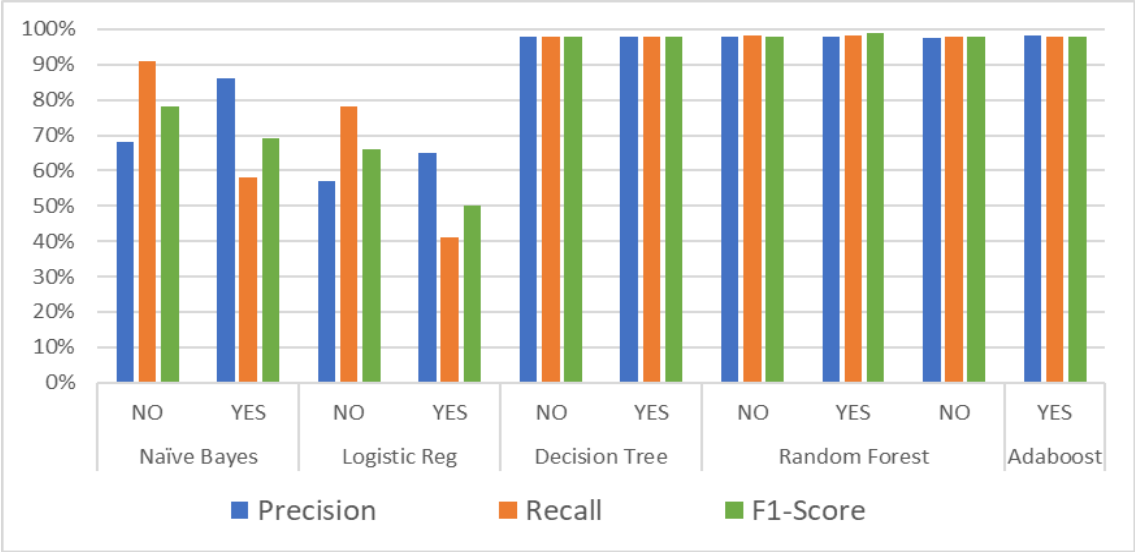

Figure 2: Results with ML models for Dataset-1

Table 3: Results with five ML models for Dataset-2

Algorithm	Class	P	R	F1	Accuracy
Naïve Bayes	NO	75%	90%	82%	79.8%
	YES	87%	70%	78%	
Logistic Reg	NO	56%	78%	66%	59.4%
	YES	65%	40%	50%	
Decision Tree	NO	98%	98%	97.70%	97.9%
	YES	98%	98%	97.80%	
Random Forest	NO	98%	98%	98%	98%
	YES	98%	98%	98%	
Adaboost	NO	97.5%	98.%	98%	98%

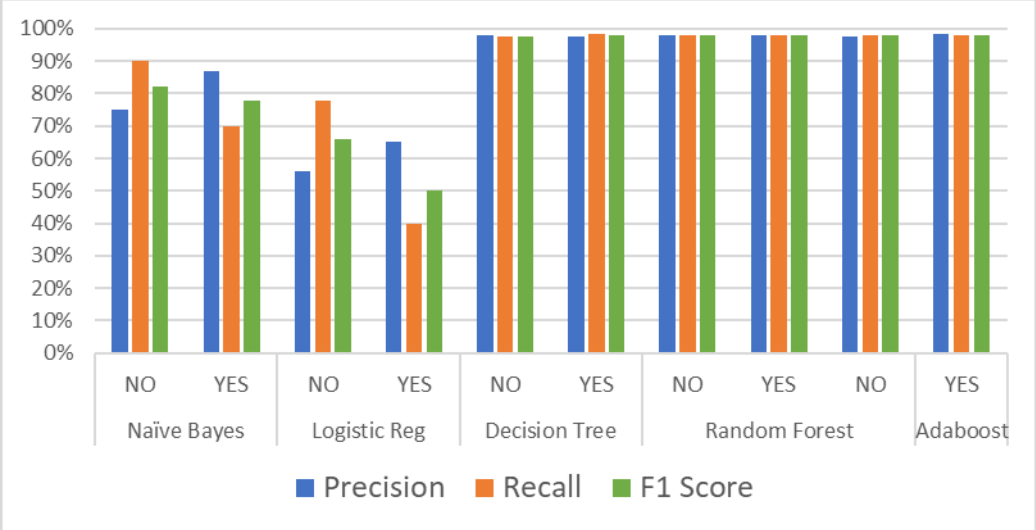

Figure 3: Results with ML models for Dataset-2

Table 4: Results with five ML models for Dataset-3

Algorithm	Class	P	R	F1	Accuracy
Naïve Bayes	NO	92%	70%	79%	81.60%
	YES	75%	93%	83%	
Logistic Reg	NO	83%	66%	74%	76.00%
	YES	71%	86%	78%	
Decision Tree	NO	98%	98%	98%	98%
	YES	98%	97.8%	97.9%	
Random Forest	NO	96%	100%	98%	98%
	YES	100%	97%	98%	
Adaboost	NO	98%	98%	98%	98%

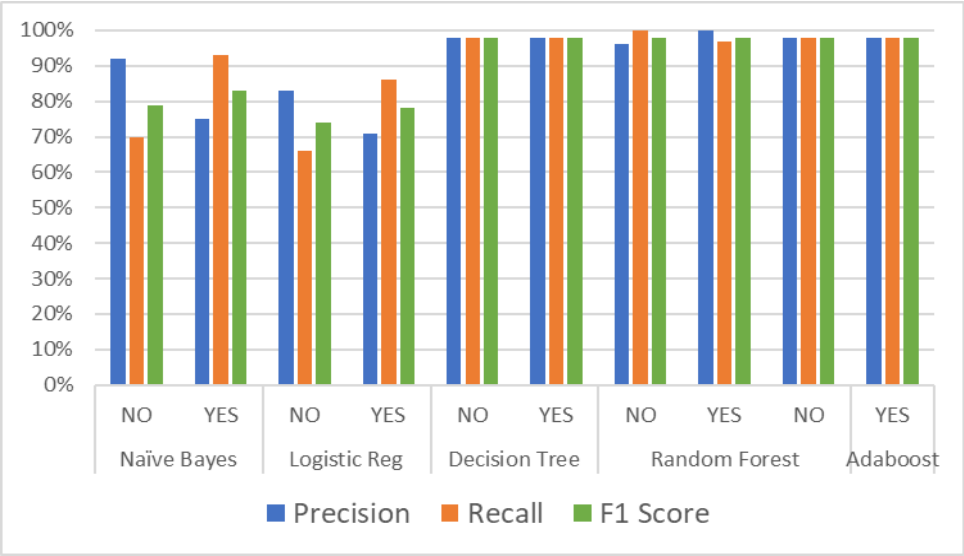

Figure 4: Results with ML models for Dataset-3

Table 5: Results with five ML models for Dataset-4

Algorithm	Class	P	R	F1	Accuracy
Naïve Bayes	NO	61%	76%	72%	73.8%
	YES	64%	68%	70%	
Logistic Reg	NO	82%	87%	85%	85%
	YES	84%	85%	86%	
Decision Tree	NO	97.8%	97.4%	96.9%	98.1%
	YES	98.3%	98%	98.7%	
Random Forest	NO	97%	99%	98%	98%
	YES	99%	98%	97.8%	
Adaboost	NO	98.1%	98%	98%	98%

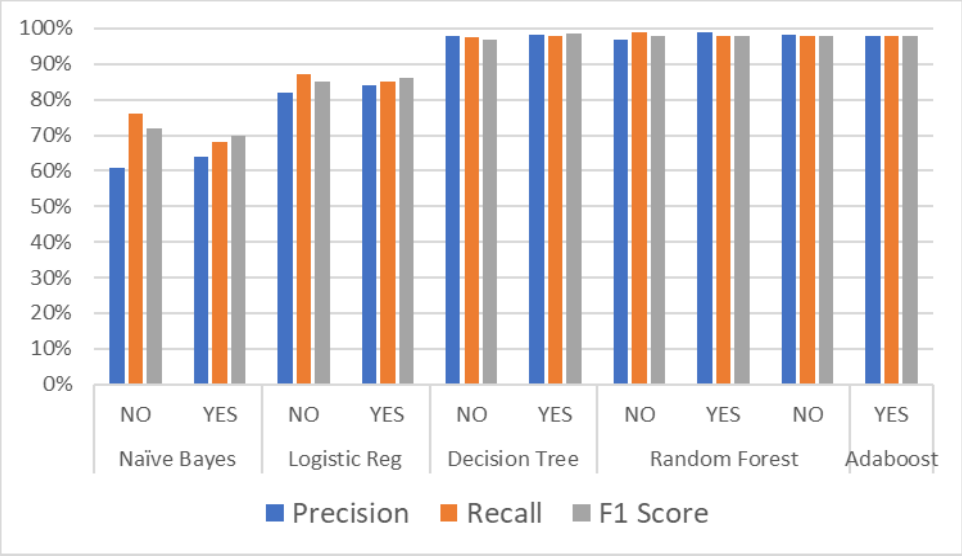

Figure 5: Results with ML models for Dataset-4

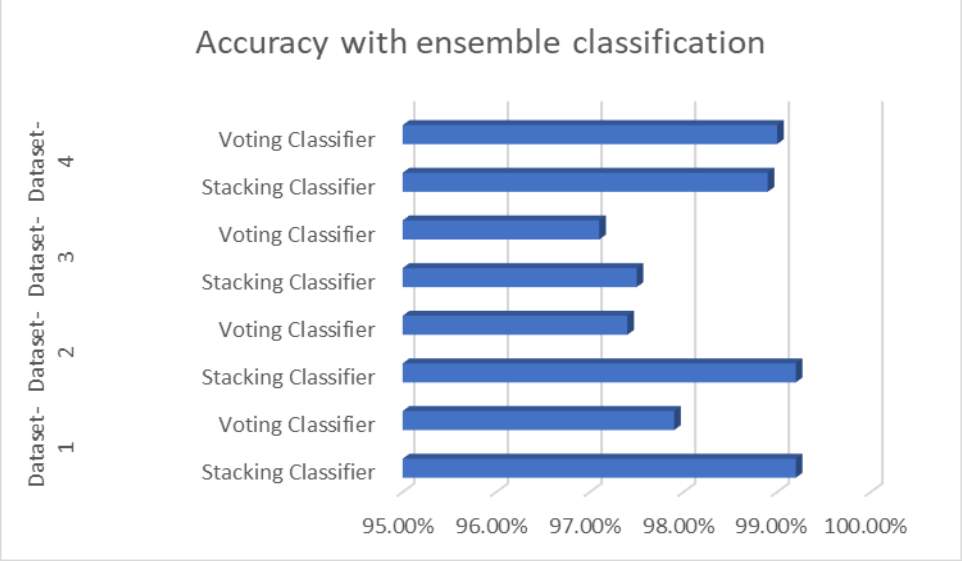

Figure 6: Accuracy with Ensemble Model

4.1.4. Apply ML algorithms with dataset-4

The proposed five algorithms are applied for Dataset-4(created through ENN technique). The results after applying algorithms are shown in Table 5. Table-5 and Figure 5 indicates that random forest, Adaboost, and decision tree yielded favorable results for insider threat detection with dataset-4. The precision, recall, and accuracy metrics all demonstrated strong performance for these three algorithms.

After applying ML model to all datasets, it is evident that Random Forest, Decision Tree, and Adaboost demonstrated superior performance across all datasets. To further enhance the model's efficacy, the proposition of an ensemble model is proposed.

4.2. Applying Ensemble Learning

The four datasets are applied with two ensemble techniques namely stacking classifier and voting classifier. As smote and adasyn datasets consists of large number of samples, only 75000 samples are used from these two datasets. For reducing complexity of the model Principal Component Analysis (PCA) applied before applying ensemble model. The results are shown in the Table 6 and Figure 6. For applying ensemble technique to over sampled datasets, only 75,000 samples are considered as the size of the dataset is very large. For both stacking and voting classifier, the combination of adaboost, RF and Decision Tree given good results.

Table 6: Accuracy with ensemble model

Algorithm	Stacking Classifier	Voting Classifier
Dataset-1	99.2%	97.9%
Dataset-2	99.2%	97.4%
Dataset-3	97.5%	97.1%
Dataset-4	98.9%	99%

4.3. Comparison with previous work

Most of the previous works for insider threat detection based on conventional ML algorithms only. The issue of data imbalance also not resolved in many cases. In this paper, the authors resolved the class imbalance issue by applying over sampling and under sampling methods. After that, ML models applied. To further increase accuracy, ensemble learning also proposed and accuracy enhanced up to 99% for all datasets created from under sampling and over sampling. So, the proposed model outperformed existing models. Table 6 shows performance of proposed work with existing models.

Table 7: Accuracy Comparison

Model	Accuracy
SVM [8]	82.4%
XGboost[4]	92%
ISOF [12]	80%
RF [14]	98.1%
Proposed Method	99%

From Table 6, it is observed that the proposed model achieved good accuracy of 99% ensemble learning. The model also resolved the issue of data imbalance. So proposed model can easily handle new datasets for insider threat detection.

5. CONCLUSION AND LIMITATIONS

The security of an organization is significantly jeopardized by insider threats. Detecting and mitigating these threats is vital for a robust cybersecurity strategy. This paper focused on utilizing ML algorithms for insider threat detection, using a well-known CERT dataset for experimentation. The initial challenge of dealing with a largely imbalanced dataset was addressed through a combination of oversampling and under sampling techniques. Three oversampling methods random oversampling, SMOTE, and ADASYN along with three under sampling techniques random under sampling, Cluster Centroids, and Edited Nearest Neighbors were employed to address data imbalance. Subsequently, five ML techniques namely Logistic Regression, Adaboost, Decision Tree, Random Forest, and Naïve Bayes were applied to datasets generated using these sampling techniques and achieved good results. To further enhance model performance, ensemble learning techniques were employed. The experimental results revealed that the proposed model outperformed existing models in the domain of insider threat detection. This highlights the effectiveness of the adopted machine learning approach and the significance of addressing data imbalance in for insider threat detection. Overall, the findings accentuate the importance of utilizing advanced techniques to bolster cybersecurity

defenses and protect organizational assets, maintain trust, and adhere to legal and regulatory requirements. The proposed method tested on a dataset CERT only. In future, this method can be tested on several insider threat datasets.

REFERENCES:

[1] Sasmita Kumari Nayak, "Classification of cyclones using machine learning techniques," World Journal of Advanced Research and Reviews, vol. 20, no. 2. GSC Online Press, pp. 433–440, Nov. 30, 2023. doi: 10.30574/wjarr.2023.20.2.2156.

[2] B. Bin Sarhan and N. Altwaijry, "Insider Threat Detection Using Machine Learning Approach," Applied Sciences, vol. 13, no. 1. MDPI AG, p. 259, Dec. 25, 2022. doi: 10.3390/app13010259.

[3] N. T. Moekthi Prajitno, H. Hadiyanto and A. F. Rochim, "Research Opportunity of Insider Threat Detection based on Machine Learning Methods," 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia, 2023, pp. 292-296, doi: 10.1109/ICAIIC57133.2023.10067010.

[4] M. Anul Haq, M. Abdul Rahim Khan, and M. Alshehri, "Insider Threat Detection Based on NLP Word Embedding and Machine Learning," Intelligent Automation Soft Computing, vol. 33, no. 1. Computers, Materials and Continua (Tech Science Press), pp. 619–635, 2022. doi: 10.32604/iasc.2022.021430.

[5] N. M. Sheykhkanloo and A. Hall, "Insider Threat Detection Using Supervised Machine Learning Algorithms on an Extremely Imbalanced Dataset," International Journal of Cyber Warfare and Terrorism, vol. 10, no. 2. IGI Global, pp. 1–26, Apr. 2020. doi: 10.4018/ijcwt.2020040101.

[6] C. Zhang, S. Wang, D. Zhan, T. Yu, T. Wang, and M. Yin, "Detecting Insider Threat from Behavioral Logs Based on Ensemble and Self-Supervised Learning," Security and Communication Networks, vol. 2021. Hindawi Limited, pp. 1–11, Nov. 26, 2021. doi: 10.1155/2021/4148441.

[7] A. Mittal and U. Garg, "Design And Analysis Of Insider Threat Detection And Prediction System Using Machine Learning Techniques," 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, 2023, pp. 1-8, doi: 10.1109/ICECCT56650.2023.10179686.

[8] A. S, S. D, and P. G, "Malicious insider threat detection using variation of sampling methods for anomaly detection in cloud environment," Computers and Electrical Engineering, vol. 105. Elsevier BV, p. 108519, Jan. 2023. doi: 10.1016/j.compeleceng.2022.108519.

[9] T. Hu, W. Niu, X. Zhang, X. Liu, J. Lu, and Y. Liu, "An Insider Threat Detection Approach Based on Mouse Dynamics and Deep Learning," Security and Communication Networks, vol. 2019. Hindawi Limited, pp. 1–12, Feb. 17, 2019. doi: 10.1155/2019/3898951.

[10] S. Yuan and X. Wu, "Deep learning for insider threat detection: Review, challenges and opportunities," Computers & Security, vol. 104. Elsevier BV, p. 102221, May 2021. doi: 10.1016/j.cose.2021.102221.

[11] R. Yousef, M. Jazzar, A. Eleyan and T. Bejaoui, "A Machine Learning Framework & Development for Insider Cyber-crime Threats Detection," 2023 International Conference on Smart Applications, Communications and Networking (SmartNets), Istanbul, Turkiye, 2023, pp. 1-6, doi: 10.1109/SmartNets58706.2023.10215718.

[12] R. B. Peccatiello, J. J. C. Gondim and L. P. F. Garcia, "Applying One-Class Algorithms for Data Stream-Based Insider Threat Detection," in IEEE Access, vol. 11, pp. 70560-70573, 2023, doi: 10.1109/ACCESS.2023.3293825.

[13] L. Liu, C. Chen, J. Zhang, O. De Vel and Y. Xiang, "Insider Threat Identification Using the Simultaneous Neural Learning of Multi-Source Logs," in IEEE Access, vol. 7, pp. 183162-183176, 2019, doi: 10.1109/ACCESS.2019.2957055.

[14] M. Fatima, O. Rehman, and I. Rahman, "Impact of Features Reduction on Machine Learning Based Intrusion Detection Systems," ICST Transactions on Scalable Information Systems. European Alliance for Innovation n.o., p. 447, Jul. 13, 2018. doi: 10.4108/eetsis.vi.447.

[15] N. Dixit, R. Gupta and P. Yadav, "Insider Threat Classification Using KNN Machine-Learning Technique," 2023 IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India, 2023, pp. 1-5, doi: 10.1109/InC457730.2023.10263010.

[16] AM. S. Sarma, Y. Srinivas, M. Abhiram, L. Ullala, M. S. Prasanthi and J. R. Rao, "Insider Threat Detection with Face Recognition and KNN User Classification," 2017 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Bangalore, India, 2017, pp. 39-44, doi: 10.1109/CCEM.2017.16.

[17] D. C. Le, N. Zincir-Heywood and M. I. Heywood, "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, March 2020, doi: 10.1109/TNSM.2020.2967721.

[18] M. Aldairi, L. Karimi and J. Joshi, "A Trust Aware Unsupervised Learning Approach for Insider Threat Detection," 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), Los Angeles, CA, USA, 2019, pp. 89-98, doi: 10.1109/IRI.2019.00027..

[19] R. Yousef, M. Jazzar, A. Eleyan and T. Bejaoui, "A Machine Learning Framework & Development for Insider Cyber-crime Threats Detection," 2023 International Conference on Smart Applications, Communications and Networking (SmartNets), Istanbul, Turkiye, 2023, pp. 1-6, doi: 10.1109/SmartNets58706.2023.10215718.

[20] G. Padmavathi, D. Shanmugapriya and S. Asha, "A Framework to Detect the Malicious Insider Threat in Cloud Environment using Supervised Learning Methods," 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2022, pp. 354-358, doi: 10.23919/INDIACom54597.2022.9763205.

[21] M. Singh, B. M. Mehtre and S. Sangeetha, "User Behavior Profiling using Ensemble Approach for Insider Threat Detection," 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA), Hyderabad, India, 2019, pp. 1-8, doi: 10.1109/ISBA.2019.8778466.

[22] https://web.cs.dal.ca/~lcd/data/CERTr5.2/

[23] Sasmita Kumari Nayak, “Analysis and High Accuracy Prediction of Coconut Crop Yield Production Based on Principle Component Analysis with Machine learning Models”, IJMA, vol. 9, no. 4, pp. 359 - 369, Dec. 2020.