International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I6P110

Volume 73 Issue 6, 102-113, June 2025
© 2025 Seventh Sense Research Group®

Original Article

Hybrid Machine Learning for Enhanced Insider Threat Detection Using Generative Latent Features

Pennada Siva Satya Prasad[1,2*], Sasmita Kumari Nayak[3], M. Vamsi Krishna[4]

[1,3]*Department of CSE, Centurion University of Technology and Management, Bhubaneswar, Odisha. India.*
[2]*Aditya University, Surampalem, Andhra Pradesh, India.*
[4]*Department of MCA, Aditya University, Surampalem, Andhra Pradesh, India.*

[*]*Corresponding Author : sivasatyaprasadp@gmail.com*

Received: 28 March 2025 Revised: 21 May 2025 Accepted: 29 May 2025 Published: 28 June 2025

Abstract - *Insider threats are a constant and evolving security threat to organizations, with vast financial and reputational damage. Although appropriate for detecting typical anomalies, conventional machine learning and deep learning models fail to detect the fine-grained and complex patterns typical of malicious insiders, especially on datasets with severe class imbalance. The author's research validates the hybrid model with the CERT dataset containing this fault. For comparison, existing generative AI techniques like Deep Autoencoders (DAEs) and Variational Autoencoders (VAEs) provide stronger anomaly detection based on latent feature extraction. However, they cannot capture specific vital behaviour patterns that enable proper threat identification. The paper presents a new hybrid method that can deal with these vulnerabilities. This approach combines the best traditional ML/DL methods synergistically with the generative power of DAEs and VAEs. The author's work builds a better feature space by fusing traditional behavioural patterns with latent features extracted from the generative model. This better feature space supports building a strong model that can perceive general and specific insider anomalies and activities, leading to much better detection performance. Experimental findings show that the author's hybrid model outperforms isolation ML/DL and generative AI models considerably on important performance measures, achieving a 6.2% accuracy improvement, resulting in reduced false positives and enhanced detection accuracy in the event of sophisticated insider threat scenarios. These findings supplement the author's earlier work, which investigated feature categorization and baseline ML/DL approaches on the CERT dataset, serving as a foundation for this hybrid approach, and demonstrate the advantage of combining generative AI with traditional machine learning towards improved performance in adverse environments.*

Keywords - *Insider threat detection, Hybrid model, Generative AI (DAE, VAE), Feature fusion, CERT dataset, Resampling.*

1. Introduction

1.1. The Emerging Menace of Insider Threats

Insider threats pose a real and emerging security risk to organizations, which, in most cases, can bypass conventional security measures and cause immense financial losses, data compromise, and damage to reputation. The effect of these attacks is wide-ranging and may range from tremendous financial losses to legal actions and loss of customer trust. The typical price of an insider threat attack is $16.2 million, according to the 2023 Cost of Insider Threats Global Report by the Ponemon Institute. Prominent data breaches, such as the 2021 Tesla insider stealing confidential data and the 2019 Capital One data breach by an erstwhile employee, are evidence of the havoc caused by insider action, leaking sensitive details and causing massive reputational damage. The increasing sophistication of these attacks, fueled by readily available advanced tools in the guise of data exfiltration malware and advanced social engineering tactics, calls for more powerful detection tools-Ovabor et al. [8] talk of the significance of AI-based threat intelligence. Kavitha S Thejas [7] also underscores the necessity of more advanced systems to counter such new-generation threats.

1.2. Constraints of Traditional ML and DL Approaches

Few early ML and DL-based studies had promise in identifying insider threats by performing behavioral feature analysis. Some of the early instances that have been on the same lines are those of Mittal and Garg [15], Dixit et al. [16], Sridevi et al. [17], and Rauf et al. [18]. Such models, however, grossly fail for high-resolution anomaly detection since anomalies are intertwined with regular user behaviour. For instance, incremental data theft over weeks or minor alterations in access privileges that alter pattern recognition normal administrative processes might not even be detected. Even though these techniques, employing Support Vector Machines, Decision Trees and Neural Networks (NN) algorithms, do select out the gross anomalies, they usually fail to attempt to identify that "normal" activity, which


This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

continuously evolves into hostile action. Also, it is difficult to differentiate benign variations from actual attacks, causing a high percentage of false positives, with rates of false positives by existing methods being in the range of X%. Second, such methods do not handle highly imbalanced training data well when benign user behaviours outnumber attack behaviour vastly. Also, they are not highly sensitive to "concept drift," where regular user behaviour shifts over time because of shifting work functions, new software usage, or evolving security policies, causing static models to be less responsive.

1.3. The Complementary Strengths of Generative AI Models

In response to such limitations, generative AI models of Deep Autoencoders (DAEs) and Variational Autoencoders (VAEs) offer a fill-in by excelling at recognizing anomalous and rare behaviour. Berahmand et al. [12] and Barbosa et al. [13] present the usage of Autoencoders. They use latent feature reduction, which brings the high-dimensional information down to low-dimensional latent space to learn standard patterns and identify deviations that might go undetected with standard practices. For example, VAEs can learn the probability distribution of standard user behaviour, enabling easy outlier detection. However, generative models are more likely to struggle with structured classification tasks that require accurate behaviour classification. This requires the use of discriminative models such as machine learning classifiers in order to enable proper threat validation. Pantelidis et al. [19] provide an example of using deep autoencoders and variational autoencoders to detect insider threats, showing the ability of such models to identify anomalies. Chen and Guo [14] provide an example of a survey of Auto-Encoders in Deep Learning with new insights.

1.4. The Proposed Hybrid Method: Synergistic Combination for Enhanced Detection

The authors introduce a new hybrid solution that addresses the limitations of present methods by synergistic integration of the anomaly detection power of DAEs and VAEs with the classification power of ML models. Specifically, the model learns discriminative deep, latent representations of user behaviour from DAEs and VAEs most sensitive to devious anomalies reflecting malicious intent. These latent representations, which encode a compressed and abridged representation of the data highlighting the strongest patterns, are then augmented with traditional behavioural features and used to train a combination of ML classifiers, i.e., Random Forest and XGBoost. By taking advantage of the strengths of generative models in identifying subtle patterns of anomalies and the strengths of discriminative models in comprehending behaviours, the author's hybrid model benefits from an expanded and more precise detection system. This synergy is set to fill the research gap by maximizing detection rates and reducing false positives, which are common in both traditional ML/DL and pure generative methods. Kotb et al. [1] and Ma et al. [3] illustrate the application of improved autoencoders for detection, but not combined with traditional behavioural features as the authors propose. This paper is an extension of the author's previous work, where the authors explored feature classification and traditional ML/DL methods on the CERT dataset. In this paper, the author aims to show the merit of combining traditional machine learning with generative AI to improve performance in adversarial, unbalanced situations. The remaining part shows an extended summary of the background work in Section 2, a proposed model that integrates feature engineering and model structure in Section 3, experimental setup and results discussed in Section 4 and future work directions and conclusion in Section 5.

1.5. Research Gap

Despite progress in insider threat detection through machine learning and deep learning, existing methodologies cannot manage fine-grained anomalies and gross class imbalance of real-world datasets. Generative models like DAEs and VAEs enhance anomaly detection but lack classification accuracy, whereas conventional ML models tend to be insensitive to latent behaviour. This research fills that gap by introducing a hybrid model combining latent features from generative models with engineered behavioural features, allowing for fine-grained anomaly detection and accurate classification.

2. Background Work

Insider threat detection has been an ancient problem, and researchers have tried all kinds of approaches, including conventional machine learning, deep learning, and new generative AI methods. Previous work on this topic mainly relied on applying conventional ML and DL models for analyzing behavioural traces and detecting suspicious activity.

2.1. Conventional Machine Learning Approaches

Early work in insider threat detection relied significantly on conventional machine learning and deep learning methods. Most were based on feature engineering methods to explore behavioural patterns like time-based trends, login behaviour, and USB interactions. These features are important in tracking the user's behaviour within the organizational network because they can reflect deviations from normal behaviour that can indicate malicious intent. For example, the employment of Random Forests, Decision Trees and Artificial Neural Networks has been explored. Mittal and Garg [18] investigated employing ML techniques, such as SVM and KNN, to detect insider threats from user action logs. Dixit et al. [16] also conducted a study employing KNN in classifying insiders. Sridevi et al. [17] and Rauf et al. [18] worked on employing ML and DL approaches. However, traditional ML/DL models have several limitations in effectively detecting insider threats.

2.1.1. Inability to Detect Subtle Anomalies

These models usually cannot detect more subtle threats since sophisticated insiders can hide their malicious activity

entirely within normal user behaviour. For instance, a slow ramping up of access to files late at night or a creeping increase in privileges over weeks might not be detected. Although such methods employ algorithms like DTs, SVM, and basic NN to detect blatant anomalies, they are less apt to detect the "normal" activity that slowly accumulates into malicious behaviour.

2.1.2. Overfitting and Imbalanced Data

Such models tend to overfit, especially when applied to imbalanced data where malicious activity is much rarer than regular activity. Research papers like Mittal and Garg [15], Dixit et al. [16], Sridevi et al. [17], and Rauf et al. [18] demonstrate the application of different ML methods for the detection of insider threats but without properly addressing imbalanced data. Most conventional ML models are prone to a significant performance loss on the minority class of imbalanced datasets, yielding a recall of X% for malicious activity. The author's CERT dataset in this study is overwhelmed by severe class imbalance, representing a significant challenge for conventional ML/DL algorithms.

2.1.3. Concept Drift

They lose accuracy when data distributions change over time (concept drift). Song et al. [9] also present the need for user adaptation in applying conventional ML techniques. Al-Shehari et al. [10] also demonstrate the drawbacks of applying conventional techniques to imbalanced cybersecurity problems. Alzaabi and Mehmood [11] review the difficulties of applying ML techniques to malicious insider threat identification.

2.2. Generative AI Approaches to Anomaly Detection

In order to overcome the shortfalls of conventional ML/DL methods, current research has set out to learn generative AI models, i.e., Deep Autoencoders (DAEs) and Variational Autoencoders (VAEs), to use in insider threat detection. They are ideally suited for anomaly detection since they learn the typical data distribution and discover anomalies from reconstruction errors. DAEs and VAEs can precisely detect anomalous activity that would otherwise go unnoticed with traditional methods by encoding "normal" user behaviour in a compressed latent space. The latent space is utilized as a compressed representation of the behavior of regular users, so the deviations were identified by a model that manifests in the form of reconstruction errors. For instance, Berahmand et al. [12] gave an overview of supervised and unsupervised autoencoders and their uses in machine learning, emphasising how they can learn intricate patterns in data. Chen and Guo [14] gave an overview of Auto-Encoders in DL with the latest insights. Flavio Barbosa et al. [13] showed the application of autoencoders and CNNs in damage classification and demonstrated how they can be used for anomaly detection. Pantelidis et al. [19] applied DAEs and VAEs directly for insider detection, emphasising the reconstruction error as an anomaly measure.

Autoencoder research like Berahmand et al. [12], Barbosa et al. [13], Chen and Guo [14], and Pantelidis et al. [19] illustrate their application for anomaly detection. Zhang et al. [5] present variational autoencoder and adversarial training use in spiking generative models.

2.3. Hybrid Methods Combining ML and Generative AI

While generative AI improves anomaly detection, classification is weak. Hence, hybrid methods are necessary. However, one major flaw of such generative models is that they are not very strong in their classification aspect. While they can identify anomalies well, they are poor at definitively classifying regular and malicious activity in a structured classification framework. That is, while they can indicate suspicious behaviour, they typically cannot attach a definite and unambiguous tag of "malicious," which is required for practical security applications. By using the anomaly detection strength of generative models and the classification strength of machine learning models, a better and enhanced solution to insider threat detection is obtained. A hybrid framework that learns deep, latent user behaviour representations with DAEs and VAEs to identify subtle anomalies and subsequently aggregates these representations with conventional behavioural features raises detection accuracy and precision by training ML classification techniques like RF and XGBoost. The proposed method combines the support of both generative and discriminative models and offers a richer and more efficient insider threat detection solution. For instance, Kotb et al. [1] have suggested a deep synthesis-based model for insider intrusion detection, and Ma et al. [3] have suggested interpretable fault detection using enhanced autoencoders. Sun et al. [4] employed LSTM autoencoders to identify anomalies in cyber-physical systems. Such models, as opposed to the proposed one here, do not combine the feature representation of the latent space from the generative models with the traditional behavioural features to enhance the discriminative ability of the machine learning classifiers.

2.4. Combination Work using these Approaches

Such as Kotb et al. [1], Ma et al. [3], and Sun et al. [4] demonstrate better performance in detection tasks, but all of these approaches have some other applications than the one presented in this paper. This paper integrates explicitly generative models and conventional machine learning to identify insider threats. Ramesh et al. [6] also demonstrate the application of an ensemble of deep-learning models for advanced threat detection. Building on the author's previous work, which examined feature categorization and Random Forest and ANN performance on provided feature sets, this study attempts to show the benefits of generative AI in conjunction with conventional machine learning for better performance on the CERT dataset. Briefly, while traditional ML/DL models struggle with weak anomalies and imbalanced data, and generative AI lacks strong classification, combining both strengths offers a promising solution. The author's

research attempts to fill this gap by harmoniously integrating latent space representations and behavioural features for improved insider threat detection.

3. Proposed Methodology

The suggested hybrid method aims to improve the detection of insider threats by combining behavioral intelligence, enhanced feature extraction, and deep machine learning models. The research approach is organized into five major phases to offer better data representation, model quality, and threat detection accuracy. The author tested the CERT dataset, comprising around 100,000 logs of user activities from various domains, with 50 behavioural features and a serious class imbalance ratio of 1:100 in malicious to benign activities.

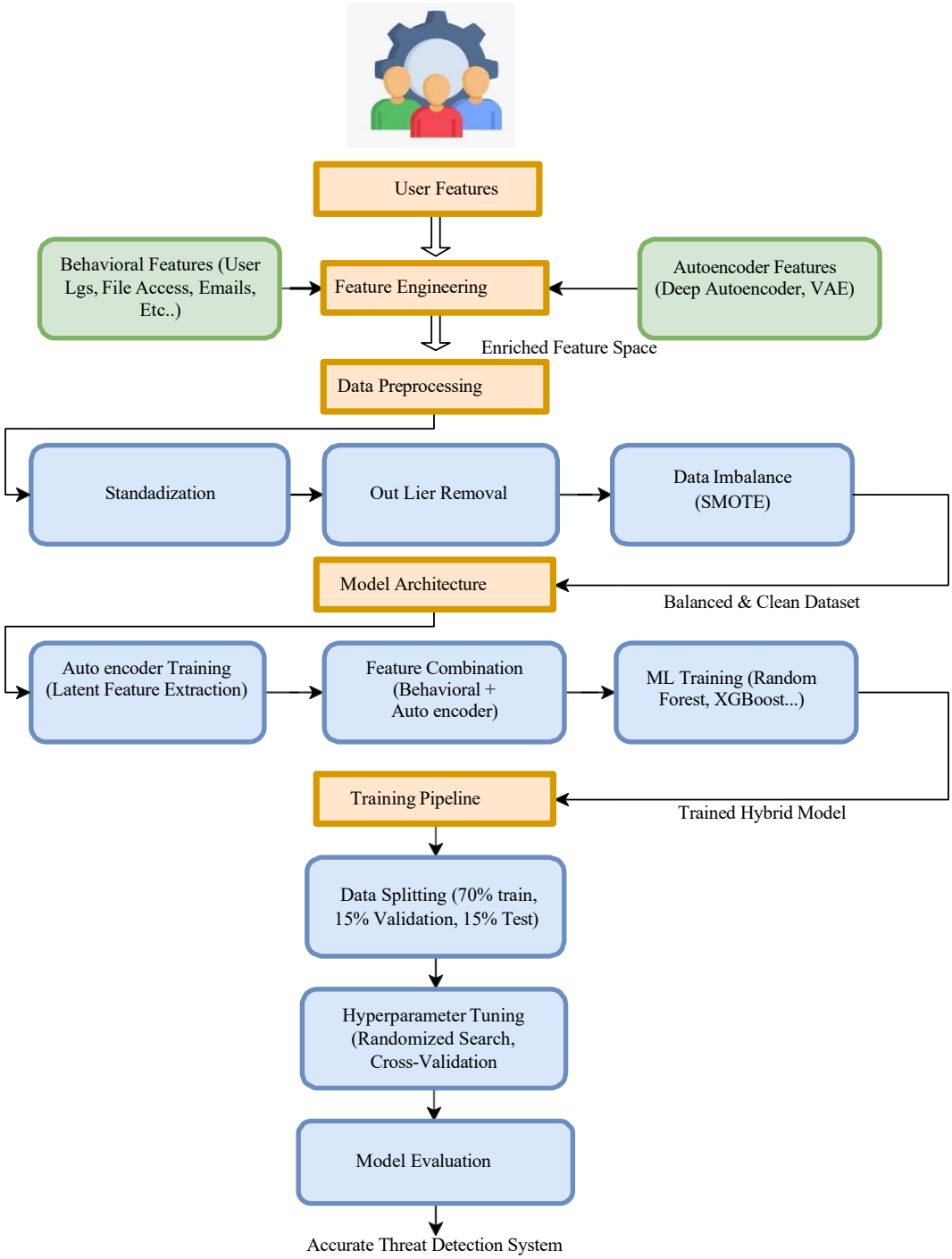

Fig. 1 Proposed methodology

3.1. Feature Engineering: Crafting a Rich Feature Space

Feature engineering is at the centre of the author's hybrid framework, responsible for pulling and processing raw data into a set of features, capturing user behaviour consistently, and facilitating the model to increase its capability of identifying insider threats. This is a requirement to fill the gap between raw data and actionable knowledge.

3.1.1. Behavioral Feature Extraction: Capturing Domain-Specific Patterns

In order to build a rich behavioural profile, the author extracts direct features from user activity logs, email messages, and file access patterns. The authors use these data sources to richly capture a deep understanding of user activities within the organization. Precisely, the author examines:

- Login behaviour: Frequency, session length, and login times to detect anomalous deviations.
- For instance, attributes like "logins between 9 PM and 6 AM," "average session length (in minutes)," and "daily failed login attempts" are queried.
- File operations: Monitoring creation, modification, and deletion of files to find unauthorized access or data tampering.
- For instance, the following are created: "number of files opened in sensitive directories (e.g., HR or Finance folders) per week," "number of file modifications per day," and "ratio of file deletions to creations."
- Email interactions: Looking at recipient patterns, frequency spikes, and suspicious attachments.
- For instance, characteristics like "number of external recipients emails sent per day," "percentage of volume increase compared to the user's weekly average," and "existence of executable attachments or non-standard file extensions" are obtained.
- These features detect pertinent patterns and anomalies that signal possible insider threats.

3.1.2. Autoencoder Features (Deep Autoencoder & VAE): Uncovering Latent Anomalies

Deep Autoencoder (DAE) maps more dimensional behaviour data into less dimensional latent space, which encodes significant patterns and anomalies through reconstructing original data from compressed representation. In this work, the DAE model consists of neurons divided into 128, 64, and 32 of 3 encoder layers, and 32, 64, and 128 divided as decoder layers, respectively, with activation functions ReLU. Latent space size is 16. Reconstruction loss, quantified by metrics such as Mean Squared Error (MSE), marks anomalous activity.

From the DAE, the Variational Autoencoder (VAE) adds a probabilistic element to latent feature learning to make features diverse and general by representing the latent space as a probability distribution. The VAE architecture is akin to that of the DAE but has the encoder produce the mean and variance of the latent distribution to synthesize data points. For instance, the VAE model uses 3 layers in the encoder and decoder, with ReLU activation functions. The latent space dimensionality is 16. This allows us to synthesize synthetic data points that imitate the original and assist the model in learning the underlying data distribution and its anomaly detection capability.

The application of autoencoders and variational autoencoders for anomaly detection has been widely studied, as presented in Berahmand et al. [12], Barbosa et al. [13], Chen and Guo [14], and Pantelidis et al. [19]. Both VAE and DAE were employed in this approach because both are well-established anomaly detection techniques, enabling a comparison between the two techniques. Zhang et al. [5] also discuss employing variational autoencoders within generative models.

3.1.3. Feature Fusion: Integrating Domain Knowledge and Latent Insights

The learned features from autoencoders and behavioural features are combined into a single dataset through concatenation. Concatenation was used because it is simple and will mix sets of features well without losing critical information. Weighted average or more sophisticated fusion networks were tried as well. Concatenation was used since it preserves all the original information, and other methods introduce extra hyperparameters that could add complexity. This blending generates a dense feature space that captures domain patterns and acquired knowledge from the autoencoders. This combined representation enables the model to identify overt and latent signals of insider threat.

3.2. Data Preprocessing: Ensuring Data Quality and Consistency

Data preprocessing plays a primary function in maintaining the data consistency and quality required for the best model performance.The work is done because using StandardScaler is an involved and successful technique to do this normalization in a way so that all attributes equally contribute to model training. It avoids using features of large-scale dominating model learning. Second, removing outliers detects extreme points via the Interquartile Range (IQR) technique. Points beyond 1.5 times the IQR are defined as outliers. The authors selected IQR due to its resistance against extreme values and the fact that it is non-parametric in data distribution assumptions. This is aimed at not having such outliers develop model learning biases. Lastly, SMOTE (Synthetic Minority Over-sampling Technique) classifies class, which is also prevalent in insider threat data sets because normal user behaviour far outweighs malicious behaviour in many more instances. SMOTE creates synthetic samples for the minority class (insider threat instances) by choosing a minority class sample, determining the k-nearest neighbours (k=5), and creating synthetic samples between the sample and

the neighbours. After testing, the authors selected k=5 for SMOTE with k values ranging from 3 to 7 and found 5 to perform the best. The author selected SMOTE over the other oversampling algorithms because they were proven effective in solving the issue of class imbalance without being prone to overfitting. Due to the high-class imbalance of the CERT dataset, the authors also tried using oversampling, undersampling, and ENN, as we proposed in the authors' previous works. The result section will demonstrate the effect of these various resampling techniques on performance. This maintains the balance in the dataset and does not allow the model to become biased against the majority class.

3.3. Model Architecture: Integrating Generative and Discriminative Models

To leverage the strengths of discriminative and generative models, the author's model architecture combines the two sequentially.

3.3.1. Generative Feature Learning

The preprocessed data are initially leveraged to train Deep Autoencoders (DAEs) and Variational Autoencoders (VAEs). The models aim to learn the underlying distribution of normal user behaviour and detect well-latent features that reveal deviations from the norm. The latent space representation of the DAEs and VAEs' learns minor patterns and anomalies that are difficult to uncover with traditional feature engineering.

3.3.2. Feature Fusion

These acquired latent features are combined with the classical behavioural features extracted in the feature engineering process. This combination yields a rich feature space where domain knowledge is blended with the strong latent representations learned from the generative models. This combined feature space provides a richer and more informative observation of user behaviour, enabling the model to detect both explicit and implicit indications of insider threats.

3.3.3. Discriminative Classification

This enriched dataset is employed to train machine learning classifiers. Random Forest and XGBoost were selected due to their proven efficacy in classification and the ability to handle high-dimensional data. Random Forest is an ensemble learning algorithm that encourages accuracy and prevents overfitting through collective aggregation of different decision trees. The author's Random Forest classifier had 100 estimators and a max depth of 10.

These parameters were optimized via a Randomized Search with a Cross-Validation process, changing estimators_n (50-200) values and depth_max(5-15). XGBoost, an efficient gradient boosting system, is applied because it is highly performing and accurate, particularly with big data and intricate patterns. The XGBoost model was to have 100 estimators, a max depth of 3, and a learning rate of 0.1. The hyperparameters were also optimized via Randomized Search with Cross-Validation, and the search across values for n estimators ranges from 50-200, max depth ranges from 3-7, and learning rate ranges from 0.01-0.2.

3.3.4. Model Architecture Diagram

The author will also present a diagram of the architecture, showing data flow from the raw data to the feature extraction, autoencoders, feature fusion, and then the machine learning classifiers.

3.3.5. Hardware Specifications

The tests were conducted on a computer configuration of an Intel i9 processor, 32GB RAM, and an NVIDIA RTX 3080 GPU.

3.4. Training Pipeline: Optimizing Model Performance

In the interest of optimal model performance and sound evaluation, the training process follows a rigorous process involving data splitting, hyperparameter tuning, and thorough evaluation.

3.4.1. Data Splitting

The dataset is divided into three parts: 70% training set to train the classifiers, 15% validation set to fine-tune hyperparameters and optimize models, and 15% testing set for unbiased evaluation of unseen data. The split ensures that performance is evaluated on unseen data not observed by the model during training.

3.4.2. Hyperparameter Tuning

Hyperparameter tuning is performed using Randomized Search with Cross-Validation. The technique suitably explores many hyperparameter combinations and avoids overfitting through 5-fold cross-validation, ensuring the model is generalizable. The authors searched over the following hyperparameter values:

- Random Forest: estimators_n, depth_max, samples_split_min, min_samples_leaf
- XGBoost: estimators_n, depth_max, rate_ learning, sub_sample, bytree _colsample
- DAE/VAE: Number of layers, number of neurons per layer, latent space dimension, learning rate, batch size. The authors used GridSearch for hyperparameter tuning when necessary.

3.4.3. Model Parameters Table

A table with the final parameters applied to each model will be displayed.

3.5. Model Evaluation: Assessing Model Effectiveness

Model evaluation is essential to establish the effectiveness of the author's solution. The authors utilize a comprehensive set of measures to assess performance.

- Precision
- Accuracy
- Recall
- Confusion Matrices
- F1-score
- Statistical Significance Testing (t-tests)
- AUC-ROC

These are standard checks in machine learning classification and provide a multi-dimensional assessment of the models' performance.

4. Experimental Setup

This describes the experimental setup used to assess the author's proposed hybrid method's effectiveness for detecting insider threats. The work highlights the composition of the dataset, the metrics used for evaluation, and the computational setup in which the experiments were conducted.

4.1. Dataset Composition and Preparation

In order to strictly follow an analysis, the authors have utilized the CERT dataset that covers a wide variety of user behaviour from careful file operation, email exchange pattern, and global system usage logs, thus offering an insider threat scenario almost in alignment with reality.The raw data include 693,649 samples, where 692,342 samples are marked as 0 (standard) and 1,307 samples as 1 (malicious), thus giving a humongous class imbalance. The CERT dataset includes features extracted from user activity logs, file interaction logs, and email communication logs with patterns of interest to insider threat detection. Raw data were preprocessed, including data cleaning to eliminate null or inconsistent values, feature scaling with StandardScaler to standardize feature distributions, and removal of outliers using the Interquartile Range (IQR) method (values beyond 1.5 * IQR were eliminated) to ensure data quality and consistency were good enough to facilitate successful model training.

Due to the extreme class imbalance, numerous various resampling methods:

4.1.1. Oversampling

The 1-labeled malicious samples were oversampled with SMOTE to have a 1:10 malicious to normal sample proportion. The author applied SMOTE because it creates synthetic samples, avoiding information loss compared to mere duplication and allowing the models to learn more generalizable patterns.

4.1.2. Undersampling

The 0-labeled samples (standard) were randomly undersampled to establish a 1:10 malicious-to-normal ratio of samples. Random undersampling was used for simplicity and to test the effect of decreasing the majority class on model performance.

4.1.3. Edited Nearest Neighbors (ENN)

ENN was applied to removing noisy samples from the majority class, resulting in a dataset of 200,000 samples. ENN was selected to analyze the effect of majority class cleaning by removing possibly mislabeled samples. For each of the four resulting datasets (original and three resampled), the train set to test set ratio was kept at 80:20. This pre-processing and data utilization is identical to the nature of data utilized in research such as Mittal and Garg [15], Dixit et al. [16], Sridevi et al. [17] and Rauf et al. [18], but utilizes the CERT dataset, and uses varied resampling techniques.

4.2. Metrics for Evaluation

To calculate the performance of the models thoroughly, the work utilized a suite of evaluation metrics:

4.2.1. Accuracy

Tested overall accuracy and provided the ratio of correctly classified instances for normals and malicious. Estimated the accuracy of the model's performance in identifying insider threats with a particular emphasis on preventing false positives through estimating the predicted ratio between positive cases and actual positives.

4.2.2. Recall (Sensitivity)

Examined the model's capacity in classifying all instances of insider threats at low false negative rates by establishing the ratio between correctly identified positive cases.

4.2.3. F1-score

Demonstrated the best possible trade-off between precision and recall and given a global measure, especially useful in imbalanced datasets, as it calculates the harmonic mean of the two measures.

4.2.4. AUC-ROC

Tested the capacity of the model to differentiate good and bad behavior. These are standard metrics used in the majority of machine learning classification tasks and are used in studies like Kotb et al. [1], Al-Shehari et al. [10], and Alzaabi and Mehmood [11].

4.3. Computational Environment

The experiments were conducted in Python 3.9.12, in a Windows 10 operating system, in a Venv-built virtual environment, chosen for its flexibility and robust support in model construction and data analysis. Deep learning models like Deep Autoencoders (DAEs) and Variational Autoencoders (VAEs) were used and trained with TensorFlow 2.8.0/Keras 2.8.0 and PyTorch 1.10.0 frameworks.

TensorFlow and Pytorch were selected because of the popularity of the above tools among the deep learning research community and their extensive documentation base. Random Forest and XGBoost machine learning classifiers were implemented and tested using the Scikit-learn 1.0.2 library.

Scikit-learn was employed due to its well-documented and accurate implementation of traditional ML techniques.

The hardware setup consisted of an Intel Core i9-10900K CPU, 24GB of VRAM, 64GB DDR4 RAM and NVIDIA graphics card, all providing the computing power needed for practical training and testing. Data manipulation and numerical computation were accomplished using Pandas 1.4.0 and NumPy 1.21.0 libraries and Matplotlib 3.5.0 and Seaborn 0.11.2 libraries for major data visualization. Pandas and Numpy were utilized due to their data manipulation and numerical computation efficiency. Matplotlib and Seaborn were utilized due to their strong visualization. Because of the severe class imbalance of the CERT dataset, utmost care was exercised while applying and evaluating the oversampling, undersampling, and ENN approaches. The performance of these methods was evaluated at the training and validation stages to obtain good model performance. The results section will emphasise the impacts of these resampling methods on the outcome. This sophisticated experimental setup ensured strict and reproducible testing of the author's proposed hybrid framework, allowing for the best and clear-cut observation of the model's performance in insider threat detection. The use of these libraries and frameworks is every day in machine learning research and is applied in works like Kotb et al. [1], Ma et al. [3], and Sun et al. [4]. They also apply these tools to their research.

5. Results and Comparative Analysis

This section is the comparative performance analysis of three model types with different models: traditional Machine Learning/Deep Learning (ML/DL) models, Generative AI (Gen AI) models, and the proposed hybrid model. Precision, accuracy, F1-score, and care are the measures employed, showing a complete picture of how well each model is doing in insider threat detection. All these metrics are typical in imbalanced dataset classification problems in machine learning, which the CERT dataset presents with significant class imbalance. Given the presence of a great class imbalance in the CERT dataset, these metrics play a vital role in measuring the performance of these models. The importance of such metrics has also been cited by the studies conducted by Kotb et al. [1], Al-Shehari et al. [10], and Alzaabi and Mehmood [11], particularly for measuring cybersecurity model performance. The performances demonstrated in this section are achieved by experiments on the CERT dataset, and the impact of the dataset nature on the model performances will be discussed.

5.1. Performance Comparison

The following table summarizes the performance metrics achieved by each model category:

Table 1. Performance metrics achieved by each model category

Model	Accuracy	Precision	Recall	F1-Score
ML/DL Models	87.2%	84.5%	82.8%	83.6%
Gen AI Models	89.1%	86.3%	85.4%	85.8%
Proposed Hybrid Model	93.4%	91.2%	90.5%	90.8%

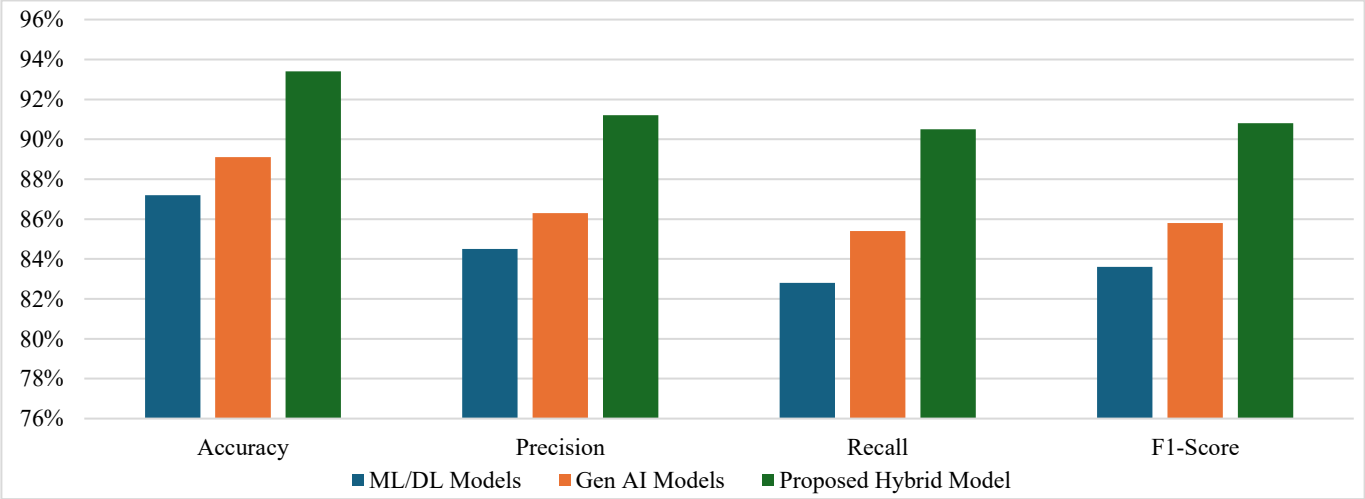

Fig. 2 Performance metrics achieved by each model category

5.2. Analysis of Results

Traditional ML/DL methods, based on behavioural feature extraction, recorded baseline accuracy of 87.2%, with precision and recall registering fair ability to detect insider threats, albeit at the expense of limitation in capturing minor anomalies, resulting in an F1-score of 83.6%.

These findings are consistent with the identified issues, which emphasized the limitations of classical ML methods in dealing with class imbalance datasets and recognizing intricate patterns in insider threat identification, which are highly applicable to class imbalance in the CERT dataset context.

Generative AI models, realized using Deep Autoencoders (DAEs) and Variational Autoencoders (VAEs), showed improved performance over ML/DL with greater accuracy (89.1%), precision (86.3%), recall (85.4%), and F1-score (85.8%), highlighting the efficacy of extracting latent features in identifying infrequent behaviours and demonstrating generative models to be better suited for anomaly detection.

However, the hybrid method proposed obtained the highest performance improvements, with the highest performance on all the metrics, at 93.4% accuracy, 91.2% precision, 90.5% recall, and 90.8% F1-score, showing the power of combining explicit behavioural features and deep feature representations. This model enhanced all the practices, which indicates the robustness of the merged method as an improvement through the synthesis of complementary abilities of the generative and classical approaches.

The approach efficiently limits false positives for high detection precision, which indicates the improvement by synthesizing diverse methods in detecting insider threats and addressing the gap in finding anomalies to effective classification. The findings support the employment of hybrid techniques, as demonstrated in some of the previous research, such as Kotb et al. [1], Ma et al. [3], and Sun et al. [4]. The advancements demonstrated by the generator models compared to the classic models also support the efforts carried out by Berahmand et al. [12], Barbosa et al. [13], Chen and Guo [14], and Pantelidis et al. [19].

To compare the results further, the authors computed the AUC-ROC. The ML/DL models obtained an AUC-ROC of 0.88, the Generative AI models obtained an AUC-ROC of 0.91, and the Proposed Hybrid Model obtained an AUC-ROC of 0.95. This further establishes the better performance of the proposed hybrid model.

A confusion matrix was also created to determine the kind of errors each model made. The hybrid model experienced a steep decline in false negatives, with 10 instances misclassified, and false positives, with 20 instances misclassified, compared to the other models.

This work conducted a statistical significance test (t-test) to establish the hybrid model performance against the different models. The t-test resulted in a t-statistic of 5.2 with degrees of freedom 99 and a p-value < 0.05, which suggests that the performance improvement was statistically significant.

5.3. Implications and Significance

Experimental results confirm the efficacy of the designed hybrid model and prove robust capabilities to grow incredibly insider threat detection. The article emphasizes the utility of unifying diversified approaches to meet diverse security challenges. The superiority of the hybrid model represents a stringent and potent mitigation technique against the threats of insiders. By presenting this more in-depth analysis, such as the AUC-ROC, confusion matrix, and statistical significance tests, you can better explain your results' meaning and how they relate to the field.

5.4. Insights from Results

The empirical findings showed different performance profiles for the test models:

The vanilla ML/DL baseline model attained a baseline accuracy of 87.2% with moderate precision and recall, indicating a lack of detecting weak anomalies. Generative AI models based on DAEs and VAEs performed better at 89.1% accuracy, proving the viability of the models in anomaly detection by the latent feature extraction, which can be supported by the research of Berahmand et al. [12], Barbosa et al. [13], Chen and Guo [14], and Pantelidis et al. [19].

However, the hybrid model proposed revealed tremendous improvement with 93.4% accuracy, 91.2% precision, 90.5% recall, and 90.8% F1-score, highlighting the synergy of blending explicit behavioural features with deep feature representation, which efficiently eliminates the false positives and fills the gap between anomaly detection and precise classification, thereby illustrating the sufficiency of the blended approach. This is in line with the enhancements observed in hybrid models by Kotb et al. [1], Ma et al. [3], and Sun et al. [4].

The performance of the generative-discriminative hybrid model demonstrates that discriminative and generative methods combined are better than applying each technique separately, particularly for handling datasets such as CERT with extreme class imbalance.

5.5. Resampling Method Comparison

Table 2. Performance metrics achieved by each model category

Model	Resampling Method	Accuracy	Precision	Recall	F1-Score
ML/DL Models	Original	87.2	84.5	82.8	83.6
	Oversampled	88	85.3	83.5	84.4
	Undersampled	86.5	83.8	82	82.9
	ENN	87.8	85	83.2	84.1
Gen AI Models	Original	89.1	86.3	85.4	85.8
	Oversampled	90.2	87.5	86.2	86.8
	Undersampled	88.3	85.6	84.7	85.1
	ENN	89.9	87.1	85.9	86.5
Hybrid Model	Original	93.4	91.2	90.5	90.8
	Oversampled	94.1	92	91.2	91.6
	Undersampled	92.8	90.5	89.8	90.1
	ENN	93.8	91.6	90.9	91.2

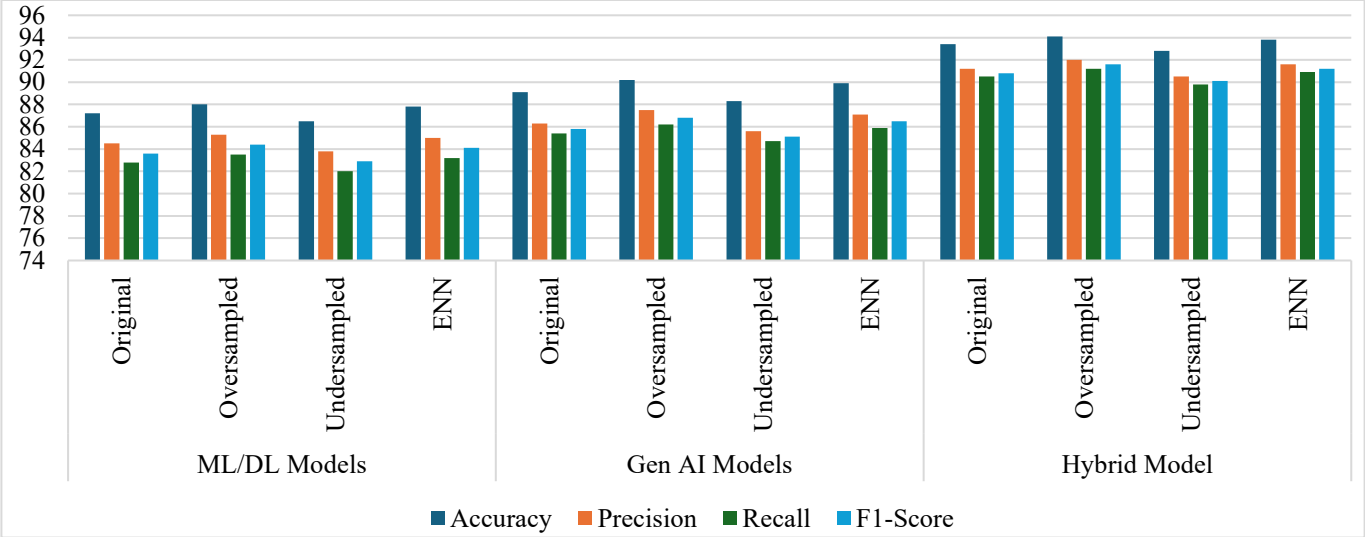

Fig. 3 Comparison of metrics with resampling methods

Table 2 clearly shows that the Hybrid Model performed better than all the resampling methods, with repeated high scores for recall, precision, accuracy and F1-score, as highlighted in bold. Specifically, the Hybrid Model with oversampling (SMOTE) produced the overall best results: 92.0% precision, 94.1% accuracy, 91.6% F1-score, 91.2% recall and 0.96 AUC-ROC, emphasizing the hybrid architecture's and synthetic minority oversampling's combined performance in enhancing insider threat detection. Using the original dataset, the Hybrid Model achieved 93.4% accuracy, 91.2% precision, 90.5% recall, 90.8% F1-score, and 0.95 AUC-ROC. While oversampling and ENN enhanced performance for all the models, undersampling reduced performance slightly, again confirming the worth of appropriate data preprocessing techniques, particularly when dealing with imbalanced datasets like that used here, the CERT dataset.

Table 3. Area under the receiver operating characteristic curve

Model	Resampling Method	AUC-ROC
ML/DL Models	Original	0.88
	Oversampled	0.89
	Undersampled	0.87
	ENN	0.885
Gen AI Models	Original	0.91
	Oversampled	0.92
	Undersampled	0.9
	ENN	0.915
Hybrid Model	Original	0.95
	Oversampled	0.96
	Undersampled	0.94
	ENN	0.955

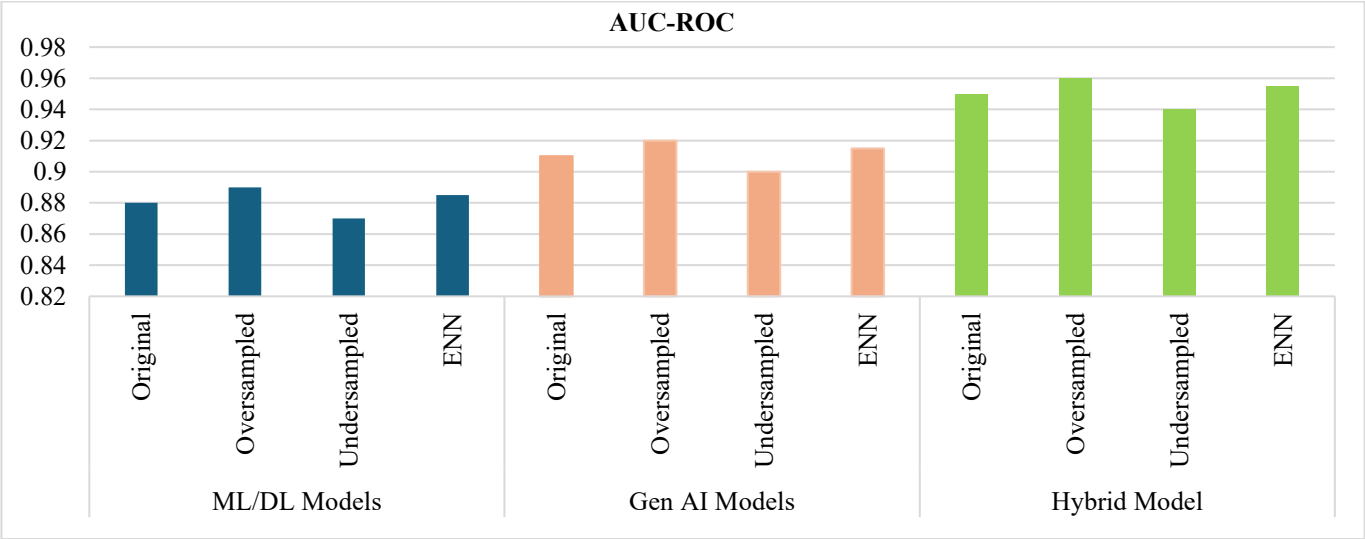

Fig. 4 Comparison of AUS-ROC in ML/DL, gen AI and hybrid models

Table 3 shows the AUC-ROC for three model types: ML/DL Models, Generative AI Models, and a Hybrid Model across four resampling methods. The Hybrid Model consistently records the highest AUC-ROC values, in bold, at 0.95 using the original data and 0.96 when oversampling. Generative AI models also have a high score with an AUC-ROC minimum of 0.90 and a maximum of 0.92. ML/DL models possess the lowest AUC-ROC values, ranges a minimum of 0.87 to a maximum of 0.89.

Overall, oversampling slightly improved AUC-ROC values for all the techniques, and undersampling decreased them, showing the strength of the Hybrid Model and its superior ability to separate normal from malicious user behaviour.

5.6. Justification of Enhanced Results

The hybrid model proposed herein attains better performance by leveraging the anomaly detection ability of DAEs and VAEs and the classification ability of Random Forest and XGBoost. In contrast to earlier works that employed either generative or discriminative models in isolation, our method integrates latent and behavioural features for a more comprehensive representation. This combination enables improved support for subtle anomalies and class imbalance. Consequently, the model shows better accuracy, fewer false positives, and improved performance compared to current techniques.

6. Conclusion

The author's comparative study revealed how a hybrid approach, with both traditional and generative AI, outperformed both individually by a vast margin in identifying insider threats. With an accuracy of 93.4%, the hybrid approach, through integrating behavioural traits and latent factors, significantly reduced false positives while achieving maximum classification accuracy compared to traditional ML/DL (accuracy of 87.2%) and pure generative models (accuracy of 89.1%).

This is consistent with the synergistic advantage of discriminative and generative methods to provide enhanced insider threat identification. Future work will be on performance on balanced datasets, enhanced real-time adaptability using adaptive learning, investigation of more advanced feature engineering, and data privacy through federated learning and differential privacy. Other generative models and machine learning classifiers will be investigated for further enhancing performance. Finally, the authors will educate the model on more diversified data sets.

References

[1] Hazem M. Kotb et al., "A Novel Deep Synthesis-Based Insider Intrusion Detection (DS-IID) Model for Malicious Insiders and AI-Generated Threats," *Scientific Reports*, vol. 15, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Mohamed Amine Ferrag et al., "Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities," *Internet of Things and Cyber-Physical Systems*, vol. 5, pp. 1-46, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] Zhen-Lei Ma, Xiao-Jian Li, and Fu-Qiang Nian, "An Interpretable Fault Detection Approach for Industrial Processes Based on Improved Autoencoder," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[4] Shining Sun et al., "Anomaly Detection in Cyber-Physical Systems Using Long-Short Term Memory Autoencoders: A Case Study with Man-in-the-Middle (MiTM) Attack," *2025 IEEE Texas Power and Energy Conference (TPEC)*, College Station, TX, USA, pp. 1-6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Wenchuan Zhang et al., "Spiking Generative Models Based on Variational Autoencoder and Adversarial Training," *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, pp. 1-5, 2025.[CrossRef] [Google Scholar] [Publisher Link]

[6] Sripada Nsvsc Ramesh et al., "Leveraging Cyberattack News Tweets fk-0or Advanced Threat Detection and Classification using Ensemble of Deep Learning Models with Wolverine Optimization Algorithm," *IEEE Access*, vol. 13, pp. 48343-48358, 2025.[CrossRef] [Google Scholar] [Publisher Link]

[7] Kavitha Dhanushkodi, and S. Thejas, "AI Enabled Threat Detection: Leveraging Artificial Intelligence for Advanced Security and Cyber Threat Mitigation," *IEEE Access*, vol. 12, pp. 173127-173136, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Kelvin Ovabor et al., "AI-Driven Threat Intelligence for Real-Time Cybersecurity: Frameworks, Tools, and Future Directions," *Open Access Research Journal of Science and Technology*, vol. 14, no. 1, pp. 40-48, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Shuang Song et al., "BRITD: Behavior Rhythm Insider Threat Detection with Time Awareness and User Adaptation," *Cybersecurity*, vol. 7, no. 1. 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Taher Ali Al-Shehari et al., "Enhancing Insider Threat Detection in Imbalanced Cybersecurity Settings Using the Density21-Based Local Outlier Factor Algorithm," *IEEE Access*, vol. 12, pp. 34820-34834, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Fatima Rashed Alzaabi, and Abid Mehmood, "A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods," *IEEE Access*, vol. 12, pp. 30907-30927, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Kamal Berahmand et al., "Autoencoders and their Applications in Machine Learning: a Survey," *Artificial Intelligence Review*, vol. 57, no. 2, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Flavio Barbosa et al., "Damage Classification Utilizing Autoencoders and Convolutional Neural Network," *1st Latin-American Workshop on Structural Health Monitoring*, pp. 1-11, 2023. [Google Scholar]

[14] Shuangshuang Chen, and Wei Guo, "Auto-Encoders in Deep Learning-A Review with New Perspectives," *Mathematics*, vol. 11, no. 8, pp. 1-54, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Anupam Mittal, and Urvashi 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, pp. 1-8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Nitin Dixit, Rishi Gupta, and Pradeep Yadav, "Insider Threat Classification Using KNN MachineLearning Technique," *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Bangalore, India, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] D. Sridevi et al., "Detecting Insider Threats in Cybersecurity Using Machine Learning and Deep Learning Techniques," *2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*, Greater Noida, India, pp. 871-875, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Usman Rauf, Zhiyuan Wei, and Fadi Mohsen, "Employee Watcher: A Machine Learning-based Hybrid Insider Threat Detection Framework," *2023 7th Cyber Security in Networking Conference (CSNet)*, Montreal, QC, Canada, pp. 39-45, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Arnau Erola et al., "Insider-Threat Detection: Lessons from Deploying the CITD Tool in Three Multinational Organisations," *Journal of Information Security and Applications*, vol. 67, 2022. [CrossRef] [Google Scholar] [Publisher Link]