European Journal of Computer Science and Information Technology, 14(2), 1-12, 2026
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Comparative Study of Ensemble and Neural Models for Insider Threat Detection Under Class Imbalance

Kitoye Ebire Okonny[1], Eseosa Omorogiuwa[2], Matthew Ehikhamenle[2]
[1]ICT Centre, Ignatius Ajuru University of Education, Port Harcourt, Nigeria
[2]Department of Electrical and Electronics Engineering, University of Port Harcourt, Port Harcourt, Nigeria

doi: https://doi.org/10.37745/ejcsit.2013/vol14n2112 Published March 01, 2026

Citation: Okonny K.E., Omorogiuwa E., Ehikhamenle M. (2026) Comparative Study of Ensemble and Neural Models for Insider Threat Detection Under Class Imbalance, *European Journal of Computer Science and Information Technology*, 14(2), 1-12

Abstract: *Insider threats remain one of the most difficult cybersecurity risks to detect because malicious activities often originate from legitimate users operating within authorised boundaries. Machine learning techniques have increasingly been applied to insider threat detection; however, there is limited empirical evidence comparing the effectiveness of classical machine learning models and deep learning architectures on large-scale behavioural datasets under realistic class imbalance conditions. This study presents a comparative performance evaluation of ensemble machine learning and neural deep learning models for insider threat detection using a large publicly available behavioural risk dataset comprising 299,880 employee activity records. After rigorous preprocessing, feature engineering, and class balancing through controlled undersampling, four models were evaluated: Random Forest, Extreme Gradient Boosting (XGBoost), Multi-Layer Perceptron (MLP), and an Autoencoder-enhanced MLP (AE-MLP). Experimental results show that ensemble tree-based methods outperform deep neural models on tabular behavioural data, with XGBoost achieving the best overall performance (Accuracy 0.894, F1-score 0.895, ROC-AUC 0.969). Deep learning models demonstrated competitive precision but lower recall, indicating reduced sensitivity to malicious behaviour patterns. To validate model behaviour, SHAP-based global feature importance analysis was applied to the best-performing model, confirming that predictions relied on meaningful behavioural indicators, including data transfer activity, printing behaviour, access timing, and employee role characteristics. The findings suggest that for structured insider threat datasets, optimised classical ensemble models remain more effective and computationally efficient than deep neural approaches, while lightweight explainability methods can provide useful behavioural validation without heavy interpretability overhead.*
Keywords: comparative study, ensemble, neural models, insider threat detection under class imbalance

INTRODUCTION

Insider threats pose an ongoing and evolving cybersecurity risk across government agencies, financial institutions, healthcare systems, and large enterprises. Unlike external attackers,

insiders operate with legitimate credentials and authorised access, allowing malicious or negligent actions to bypass traditional perimeter-based defences (Rathore, 2025). Insider incidents may include data exfiltration, intellectual property theft, misuse of privileged access, and policy violations, often carried out through patterns of behaviour that appear normal at first glance. The detection of such threats, therefore, requires behavioural modelling approaches capable of identifying subtle deviations from established activity norms (Inayat et al., 2024; Diana et al., 2025; Kamatchi & Uma, 2025).

Recent advances in machine learning have significantly improved anomaly detection and behavioural risk modelling across cybersecurity domains. Both classical machine learning algorithms and deep learning architectures have been applied to insider threat detection tasks using activity logs, access records, and behavioural aggregates (Ofori et al., 2025; Tao et al., 2025). Ensemble tree-based methods such as Random Forest and gradient boosting have demonstrated strong performance on structured tabular datasets, while deep neural networks offer powerful representation learning capabilities that may capture complex behavioural relationships (Cha et al., 2021; Reddy et al., 2025; Inyang & Johnson, 2025; Mienye & Swart, 2024; Feng et al., 2024). However, despite widespread adoption, there remains a limited systematic comparison between classical machine learning and deep learning approaches for insider threat detection under realistic organisational data conditions.

A major practical challenge in insider threat detection is severe class imbalance. Malicious insider activities typically represent a small fraction of total user behaviour, often below five per cent in operational datasets (Zheng et al., 2021; Yi & Tian, 2024). This imbalance can bias predictive models toward benign classifications, reducing recall for malicious cases and weakening operational usefulness. Model comparison studies must therefore consider imbalance mitigation strategies, robustness, and detection sensitivity rather than accuracy alone (Javed et al., 2024; Raftopoulos et al., 2025). Furthermore, while explainable artificial intelligence techniques are increasingly integrated into security analytics, their role varies across research objectives (Sharma et al., 2025). In performance-centric comparative studies, explainability serves primarily as a validation mechanism to confirm that high-performing models rely on meaningful behavioural indicators rather than spurious correlations.

This study conducts a comparative evaluation of classical machine learning and deep learning models for insider threat detection using a large behavioural dataset derived from organisational activity indicators. The evaluated models include Random Forest, Extreme Gradient Boosting (XGBoost), Multi-Layer Perceptron (MLP), and an Autoencoder-enhanced MLP (AE-MLP). The investigation focuses on detection performance, robustness under class imbalance handling, and modelling effectiveness for structured behavioural data. Explainability is incorporated in a limited, supportive capacity through SHAP-based global feature importance analysis of the best-performing model, as this method provides reliable and transparent insights into how key behavioural features influence model predictions without shifting the primary focus away from performance evaluation (Lu et al., 2023; Attai et al., 2023). This study investigates which modelling paradigm, ensemble machine learning or deep learning, achieves superior detection performance for insider threats under realistic behavioural data conditions. The main contributions of this study are as follows:

- A comparative evaluation of ensemble machine learning and deep learning models on a large insider threat behavioural dataset
- Performance analysis using multiple evaluation metrics, including precision, recall, F1-score, and ROC-AUC
- Examination of model behaviour under class imbalance mitigation through reproducible undersampling
- Lightweight explainability validation using SHAP global feature importance to confirm behavioural relevance of model predictions

The remainder of this paper is organised as follows. Section 2 presents the methodology adopted in this study, including a detailed description of the insider threat dataset, preprocessing procedures, behavioural feature selection, data cleaning, and feature scaling using Min–Max normalization. Section 3 describes the experimental setup, outlining the evaluated machine learning and deep learning models, hyperparameter optimisation strategy, configuration of the multi-layer perceptron, the autoencoder-enhanced neural model, and the performance evaluation metrics. Section 4 reports and analyses the experimental results, providing a comparative performance assessment of ensemble and neural models for insider threat detection. The Conclusion summarises the key findings, discusses practical implications for organisational security, and highlights directions for future research.

METHODOLOGY

This study adopts a structured experimental methodology to compare the detection effectiveness of classical machine learning and deep learning models for insider threat detection using behavioural risk indicators. The methodology consists of dataset preparation, feature preprocessing and transformation, class imbalance mitigation, normalisation, model development, hyperparameter optimisation, and performance evaluation. A controlled and reproducible pipeline was implemented to ensure fair comparison across all models. The overall workflow includes data cleaning, feature encoding, temporal transformation, normalisation, class balancing through undersampling, and supervised model training. Four predictive models were evaluated: Random Forest, XGBoost, MLP, and an AE-MLP. Performance comparison focused on detection capability, robustness, and modelling effectiveness on structured behavioural data. This study adopts a structured experimental methodology to compare the detection effectiveness of classical machine learning and deep learning models for insider threat detection using behavioural risk indicators. The methodology consists of dataset preparation, feature preprocessing and transformation, class imbalance mitigation, normalisation, model development, hyperparameter optimisation, and performance evaluation as presented in Figure 1.

The diagram illustrates the end-to-end workflow adopted in this study for insider threat detection, beginning with the Kaggle insider threat behavioural dataset and progressing through structured data preparation, modelling, and decision support stages. The pipeline starts with data preparation, where cleaning, encoding, feature transformation, and missing value imputation are performed to ensure analytical consistency. The processed data then undergoes class balancing through random undersampling and MinMax normalisation to address class

imbalance and scale feature values. The normalised dataset is used to train and evaluate multiple models, whose predictive outputs are assessed using standard performance metrics. SHAP analysis is applied to the best-performing model (XGBoost) to provide global feature importance insights. The combined evaluation metrics and SHAP results support final decision-making by identifying the most effective and behaviourally meaningful model for insider threat detection.

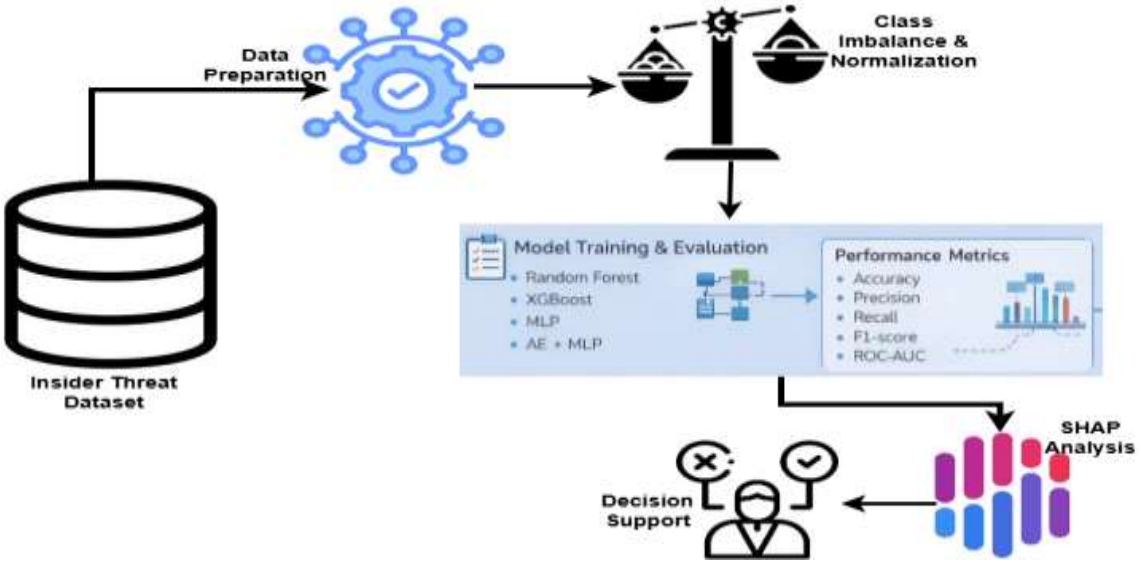

Figure 1. Workflow for Insider Threat Detection

Dataset Description and Preprocessing

This study utilises a publicly available insider threat behavioural dataset obtained from Kaggle (Bdil, 2025). The dataset is designed to model employee behavioural risk within an organisational environment using aggregated activity indicators rather than raw event logs. Each record represents a behavioural instance associated with an employee's operational activity profile. The original dataset contains 299,880 records and 48 attributes before preprocessing. The class distribution reflects realistic insider threat sparsity: 284,940 records (approximately 95%) are labelled non-malicious, while 14,940 records (approximately 5%) are labelled malicious. This imbalance mirrors real organisational conditions where insider incidents are rare relative to normal activity. Although the dataset is aggregated, it preserves implicit temporal and behavioural characteristics through summary indicators such as total presence duration, night-time access flags, early entry and late exit indicators, weekend access markers, printing behaviour, and data transfer proxies.

Feature selection focused on retaining behavioural and contextual attributes with predictive relevance to insider threat detection. Attributes exhibiting excessive missingness and limited analytical value were excluded. Specifically, the variables trip_day_number and country_name were removed due to sparsity and negligible contribution to behavioural modelling. Identifier-related fields, including employee_id and date, were also removed before model training to prevent information leakage and ensure that models learned behavioural patterns rather than individual identities. This step is critical in behavioural risk modelling to avoid artificial performance inflation.

Data cleaning involved standardising variable types and ensuring numerical consistency across all features. Boolean attributes were converted into binary numerical representations (0 and 1). Categorical attributes describing organisational role, classification level, and location context were transformed using one-hot encoding. This encoding strategy supports both tree-based algorithms and neural network models while enabling the learning of non-linear interactions between contextual and behavioural variables. Feature engineering was applied to preserve behaviourally meaningful indicators. Derived variables representing access irregularities and deviation patterns, such as off-hours activity flags, early entry indicators, late exit indicators, and weekend access markers, were retained because they are strongly associated with insider threat behaviour in prior research.

Temporal attributes such as first entry and last exit times were transformed into numerical representations expressed as minutes since midnight. This conversion enables models to capture time-of-day behavioural patterns using continuous numerical inputs. Missing values were handled using feature-specific imputation strategies guided by attribute semantics. For time-related variables, missing values were imputed using group-based median values computed from employees with similar organisational roles. This approach preserves behavioural realism better than global imputation and reduces distortion of role-dependent access patterns.

The original dataset exhibits severe class imbalance, with malicious records representing approximately five per cent of all observations. Such an imbalance can bias predictive models toward the majority class and significantly reduce malicious case detection sensitivity. To mitigate this effect, random undersampling of the majority class was applied. All malicious records (14,940) were retained, while an equal number of non-malicious records were randomly selected without replacement. This produced a balanced dataset of 29,880 records with a 1:1 class ratio. Undersampling was implemented using the resample function from sklearn.utils with a fixed random seed (random_state = 42) to ensure reproducibility. This approach was selected because it prevents majority-class dominance during training, reduces computational overhead, and is well-suited for ensemble and tree-based models commonly used in insider threat detection.

Feature scaling was performed using Min-Max normalisation, which transforms numerical features into a bounded range between 0 and 1, and the pre-processed dataset is presented in Figure 2. Normalisation was applied after encoding and feature transformation to ensure consistent feature magnitudes across variables. Min-Max scaling is particularly beneficial for neural network optimisation because it stabilises gradient updates and accelerates convergence. While tree-based models are generally scale-invariant, applying the same normalised inputs across all models ensures fairness and consistency in comparative evaluation.

employee	is_contrac	employee	has_foreig	has_crimi	has_medi	risk_trave	num_prin	total_prin	num_prin	num_prin	num_collo	num_bw_	ratio_collo	printed_fi	print_cam	num_burr	max_requ	avg_requ	num_burr
0.54749	-0.37571	-0.04827	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	0.015822	-0.41811	-0.4323	-0.18242
-0.65331	2.661657	-1.38873	-0.4774	5.027891	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	-0.39636	-0.41811	-0.4323	-0.18242
-1.32042	-0.37571	-0.04827	2.094676	-0.19889	-0.43645	-0.00818	1.381759	2.387795	-0.17227	-0.13149	2.562331	2.08092	1.959562	4.042505	2.36472	-0.39636	-0.41811	-0.4323	-0.18242
0.013802	-0.37571	1.292189	-0.4774	-0.19889	-0.43645	-0.00818	0.197236	-0.09874	-0.17227	-0.13149	-0.25516	0.001197	0.035601	-0.24737	0.863885	-0.39636	-0.41811	-0.4323	-0.18242
0.013802	-0.37571	-1.38873	-0.4774	-0.19889	2.291231	-0.00818	1.381759	0.100185	-0.17227	-0.13149	0.013176	0.143247	1.125845	-0.24737	0.863885	0.015822	-0.41811	-0.4323	-0.18242
0.013802	-0.37571	-0.04827	-0.4774	-0.19889	2.291231	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	-0.39636	-0.41811	-0.4323	-0.18242
-1.32042	-0.37571	-0.04827	-0.4774	-0.19889	-0.43645	-0.00818	0.197236	3.805119	-0.17227	-0.13149	1.556086	4.617819	0.548657	4.042505	3.865556	-0.39636	-0.41811	-0.4323	-0.18242
0.54749	-0.37571	-0.04827	-0.4774	-0.19889	-0.43645	-0.00818	0.592077	0.100185	2.375986	1.125991	-0.25516	0.285297	-0.54159	4.042505	3.865556	-0.39636	-0.41811	-0.4323	-0.18242
0.880912	2.661657	-0.04827	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	-0.39636	-0.41811	-0.4323	-0.18242
-0.92015	-0.37571	-0.04827	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	0.015822	0.559899	0.880019	-0.18242
1.481443	-0.37571	-0.04827	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	0.428006	1.537906	1.536178	1.090814
0.147224	-0.37571	-0.04827	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	-0.39636	-0.41811	-0.4323	-0.18242
0.414068	2.661657	-1.38873	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	1.252373	1.537906	1.536178	-0.18242
0.280646	-0.37571	-0.04827	2.094676	-0.19889	2.291231	-0.00818	0.986918	1.741297	-0.17227	-0.13149	-0.18807	2.593608	-0.41332	-0.24737	0.863885	0.015822	1.537906	2.192337	-0.18242
1.881709	2.661657	-0.04827	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	-0.39636	-0.41811	-0.4323	-0.18242
0.013802	-0.37571	1.292189	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	-0.39636	-0.41811	-0.4323	-0.18242
0.414068	-0.37571	-0.04827	2.094676	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	-0.39636	-0.41811	-0.4323	-0.18242
-1.58726	-0.37571	1.292189	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	-0.39636	-0.41811	-0.4323	-0.18242
-0.65331	-0.37571	1.292189	-0.4774	-0.19889	-0.43645	-0.00818	0.986918	-0.24793	-0.17227	-0.13149	-0.25516	-0.21188	0.676921	-0.24737	0.863885	-0.39636	-0.41811	-0.4323	-0.18242
0.814334	-0.37571	1.292189	-0.4774	-0.19889	-0.43645	-0.00818	-0.59245	-0.34739	-0.17227	-0.13149	-0.25516	-0.35393	-0.54159	-0.24737	-0.63695	-0.39636	-0.41811	-0.4323	-0.18242

Figure 2. Clean Insider Threats Dataset

Experimental Setup

Evaluated Models

This study evaluates four supervised learning models to enable a controlled comparison between classical machine learning and deep learning paradigms for insider threat detection. Two ensemble-based classical machine learning models and two neural network–based deep learning models were selected to reflect widely adopted approaches for structured behavioural classification tasks. The models include Random Forest and XGBoost, both of which are known for strong performance on tabular datasets and robustness to feature interactions and nonlinearity (Imani et al., 2025). The deep learning models include MLP and AE-MLP. The MLP serves as a baseline neural classifier, while the AE-MLP incorporates representation learning through dimensionality compression before classification. This selection enables a balanced and meaningful comparison between ensemble learning and neural representation learning approaches under identical data conditions.

Hyperparameter Optimisation

To ensure fair and competitive performance across all evaluated models, hyperparameter optimisation was conducted using structured grid-based search spaces tailored to each algorithm. For the Random Forest model, the tuning process explored variations in the number of trees, maximum tree depth, minimum samples required for node splitting, minimum samples per leaf, and the number of features considered at each split. These parameters directly influence model complexity, generalisation capacity, and resistance to overfitting. For XGBoost, the search space included the number of boosting estimators, maximum tree depth, learning rate, subsampling ratios for both rows and columns, and the gamma regularisation parameter controlling split quality thresholds. These parameters jointly regulate boosting strength, model regularisation, and variance control. For the neural MLP model, hyperparameter tuning covered hidden layer sizes, L2 regularisation strength, and optimisation configuration using the Adam solver with rectified linear unit activation. The use of structured hyperparameter grids ensured that each model class was evaluated under well-tuned and reproducible settings rather than default configurations.

Multi-Layer Perceptron Configuration

The Multi-Layer Perceptron model was configured as a feedforward neural classifier designed to learn nonlinear relationships among behavioural features (Abdurrakhman et al., 2025; Abu-Doush et al., 2023). The architecture search considered single hidden-layer structures with moderate neuron counts to balance representational capacity and overfitting risk. Rectified Linear Unit activation functions were used to introduce nonlinearity and support stable gradient propagation. Optimisation was performed using the Adam optimiser due to its adaptive learning rate properties and proven effectiveness in tabular neural learning tasks. Regularisation strength was varied through L2 penalty parameters to improve generalisation. A constant learning rate schedule was employed to maintain stable convergence behaviour. This configuration provides a controlled neural baseline for comparison with ensemble tree-based approaches.

Autoencoder-Enhanced Neural Model (AE-MLP)

To evaluate whether learned compressed representations improve insider threat detection, an autoencoder-enhanced neural architecture was developed. The autoencoder component was designed to learn a lower-dimensional latent representation of the normalised feature space before classification. The encoder network maps the input features through a hidden layer into a compact bottleneck representation, while the decoder reconstructs the original feature space from this compressed encoding (Wei et al., 2020; Saminathan et al., 2023). The architecture consists of an input layer followed by a 32-unit hidden layer and a 16-unit latent bottleneck layer, with a symmetric decoder structure that reconstructs the original input dimensionality. Training was performed using the Adam optimiser with a learning rate of 0.001 and mean squared error reconstruction loss. After training, the encoder portion was used to transform the original features into latent representations, which were then provided as inputs to a downstream MLP classifier. This hybrid approach allows assessment of whether representation compression and noise reduction improve classification performance for behavioural insider threat indicators.

Evaluation Metrics

Model performance was evaluated using multiple metrics to provide a comprehensive assessment of detection capability. Accuracy was used to measure overall classification correctness, while precision quantified the proportion of predicted malicious cases that were truly malicious. Recall measures the proportion of actual malicious cases correctly identified by the model, and is particularly important in insider threat detection, where missed threats carry significant operational risk. The F1-score was included as a harmonic mean of precision and recall to balance detection correctness and sensitivity. In addition, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was computed to assess class separability across decision thresholds. Using multiple metrics ensures that model comparison reflects not only overall correctness but also threat detection sensitivity and discrimination strength, which are critical in security-focused predictive modelling.

RESULTS

The comparative performance of the evaluated models, XGBoost, Random Forest, MLP, and AE-MLP, was assessed using accuracy, precision, recall, F1-score, and ROC-AUC. All models

were trained and tested under identical preprocessing, normalisation, and class-balancing conditions to ensure the fairness of comparison. The experimental results show clear performance differences between classical ensemble machine learning models and deep learning architectures on the insider threat behavioural dataset, as shown in Table 1. XGBoost achieved the strongest overall performance across nearly all evaluation metrics, with an accuracy of 0.8944, precision of 0.8888, recall of 0.9016, F1-score of 0.8952, and ROC-AUC of 0.9691. Random Forest followed as the second-best performer, achieving an accuracy of 0.8638 and ROC-AUC of 0.9458, with balanced precision and recall values. The deep learning models demonstrated lower overall detection performance. The MLP achieved an accuracy of 0.8293 and ROC-AUC of 0.9198, while the AE-MLP achieved an accuracy of 0.7877 and ROC-AUC of 0.8704. Although AE-MLP produced relatively high precision (0.8681), its recall dropped substantially to 0.6784, indicating that many malicious cases were not detected. This recall reduction is operationally significant in insider threat scenarios where missed detections are costly.

Table 1. Diagnostic Models Test Performance

Algorithm		Accuracy	Precision	Recall	F1-score	ROC-AUC
Machine Learning	**XGBOOST**	0.8944	0.8888	0.9016	0.8952	0.9691
	RF	0.8638	0.8715	0.8534	0.8624	0.9458
Deep Learning	**MLP**	0.8293	0.861	0.7855	0.8215	0.9198
	AE-MLP	0.7877	0.8681	0.6784	0.7616	0.8704

The results indicate that classical ensemble machine learning models provide superior detection effectiveness compared to deep learning models for structured insider threat behavioural data, as presented in Figure 3. XGBoost consistently achieved the best performance across accuracy, recall, F1-score, and ROC-AUC, demonstrating strong class separation capability and balanced detection behaviour. Random Forest also delivered robust results, confirming the effectiveness of bagging-based ensemble strategies for tabular behavioural risk indicators. Deep neural models showed mixed behaviour. The baseline MLP demonstrated moderate performance but did not match ensemble methods in recall or ROC-AUC. This suggests that shallow feedforward neural networks may not automatically outperform ensemble learners on engineered tabular features, even after normalisation. The AE-MLP hybrid model, which incorporated representation compression through an autoencoder, did not improve detection effectiveness and instead reduced recall significantly. This suggests that aggressive latent compression may remove subtle behavioural signals necessary for insider threat discrimination.

A key observation is that deep models achieved relatively strong precision but weaker recall compared to ensemble models. From an operational security perspective, this implies that neural models were more conservative in flagging threats but missed more malicious instances. In contrast, XGBoost achieved both high precision and high recall, making it more suitable for detection-oriented deployment scenarios. These findings are consistent with broader tabular learning research, which has repeatedly shown that gradient boosting and ensemble trees often outperform deep neural networks on structured datasets unless extremely large-scale or raw-feature inputs are available.

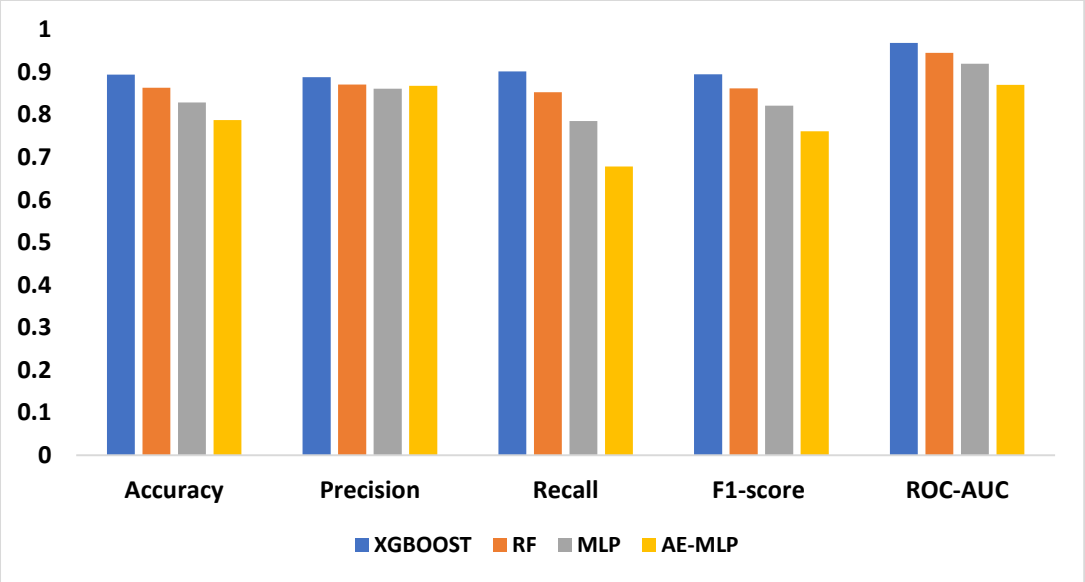

Figure 3. Model Test Performance

To validate whether the best-performing model relied on behaviourally meaningful predictors rather than spurious correlations, SHAP global feature importance analysis was applied to the XGBoost model as presented in Figure 4. Only a global importance interpretation was performed, consistent with the study's performance-focused objective and lightweight explainability scope. The SHAP summary results indicate that behavioural and contextual risk indicators primarily drive the model's predictions. The most influential features include employee seniority duration, total files burned to media, employee classification level, printing activity counts, number of black-and-white prints, number of unique campus accesses, foreign citizenship status, entry frequency, total printed pages, and contractor status. Additional influential indicators include entry and exit timing variables, weekend access flags, total presence duration, and request classification levels. These features align closely with known insider threat behavioural risk factors, including abnormal data transfer behaviour, unusual printing volume, irregular access patterns, privilege-level context, and organisational role characteristics. The SHAP results, therefore, provide supporting evidence that the high-performing model is learning behaviourally meaningful risk patterns rather than artefacts of preprocessing.

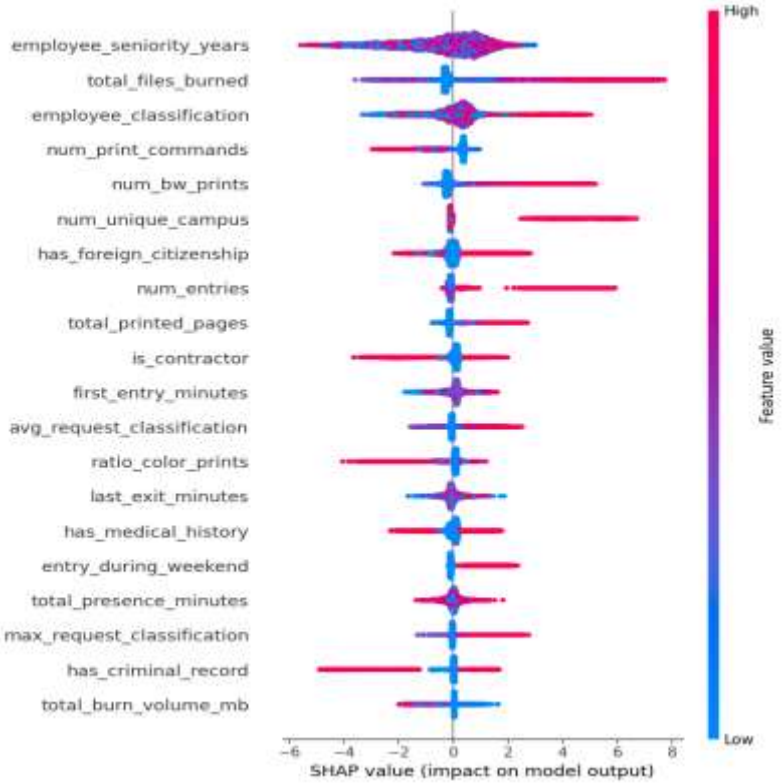

Figure 4: SHAP Summary Plot for XGBoost

CONCLUSION

This study presented a controlled comparative evaluation of classical machine learning and deep learning models for insider threat detection using a large behavioural risk dataset and a reproducible preprocessing pipeline. The investigation focused on detection performance, robustness under imbalance mitigation, and modelling effectiveness for structured organisational behavioural indicators. The results show that ensemble machine learning methods, particularly XGBoost, outperform deep neural approaches across accuracy, recall, F1-score, and ROC-AUC. Random Forest also demonstrated strong and stable performance. Deep learning models achieved competitive precision but lower recall, indicating weaker sensitivity to malicious behaviour patterns. The Autoencoder-enhanced neural model did not improve detection performance and reduced recall due to representation compression effects. Lightweight SHAP global feature analysis confirmed that the best-performing model relied on behaviourally meaningful risk indicators, including access irregularities, data transfer proxies, printing behaviour, and organisational role context. This supports the behavioural validity of the learned decision patterns. Future research can extend this work by evaluating hybrid ensemble–neural architectures that combine boosting with learned embeddings, as well as testing cost-sensitive and semi-supervised learning approaches that better reflect real-world insider threat prevalence. Additional investigation using sequential event logs and temporal deep learning models may further clarify the conditions under which deep architectures outperform ensemble methods. Incorporating streaming detection settings and cross-organisational validation datasets would also strengthen generalizability. Finally, expanded

explainability analysis integrating temporal attribution and policy-aligned explanations could improve operational trust and analyst adoption.

REFERENCES

Abdurrakhman, A., Sutiarso, L., Ainuri, M., Ushada, M., & Islam, M. P. (2025). A Multilayer Perceptron Feedforward Neural Network and Particle Swarm Optimization Algorithm for Optimizing Biogas Production. Energies, 18(4), 1002. https://doi.org/10.3390/en18041002

Abu-Doush, I., Ahmed, B., Awadallah, M. A., Al-Betar, M. A., & Rababaah, A. R. (2023). Enhancing multilayer perceptron neural network using archive-based harris hawks optimizer to predict gold prices. Journal of King Saud University-Computer and Information Sciences, 35(5), 101557. https://doi.org/10.1016/j.jksuci.2023.101557

Attai, K., Akwaowo, C., Asuquo, D., Esubok, N. E., Nelson, U. A., Dan, E., ... & Uzoka, F. M. (2023, December). Explainable AI modelling of comorbidity in pregnant women and children with tropical febrile conditions. In International Conference on Artificial Intelligence and its Applications (pp. 152-159).

Bdil, E. (2025). Insider Threat Dataset for Classified Environments [Dataset]. Kaggle. https://www.kaggle.com/datasets/efchbd1013/insider-threat-dataset-for-classified-environme

Cha, G.-W., Moon, H.-J., & Kim, Y.-C. (2021). Comparison of Random Forest and Gradient Boosting Machine Models for Predicting Demolition Waste Based on Small Datasets and Categorical Variables. International Journal of Environmental Research and Public Health, 18(16), 8530. https://doi.org/10.3390/ijerph18168530

Diana, L., Dini, P., & Paolini, D. (2025). Overview on Intrusion Detection Systems for Computers Networking Security. Computers, 14(3), 87. https://doi.org/10.3390/computers14030087

Feng, S., Yao, R., Hess, S., Daziano, R. A., Brathwaite, T., Walker, J., & Wang, S. (2024). Deep neural networks for choice analysis: Enhancing behavioral regularity with gradient regularization. Transportation Research Part C: Emerging Technologies, 166. https://doi.org/10.1016/j.trc.2024.104767

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. Technologies, 13(3), 88. https://doi.org/10.3390/technologies13030088

Inayat, U., Farzan, M., Mahmood, S., Zia, M. F., Hussain, S., & Pallonetto, F. (2024). Insider threat mitigation: Systematic literature review. Ain Shams Engineering Journal, 15(12). https://doi.org/10.1016/j.asej.2024.103068

Inyang, U. P., & Johnson, E. A. (2025). Performance comparison of XG-Boost and Random Forest for the prediction of students' academic performance. European Journal of Computer Science and Information Technology, 13(2), 1-21. https://doi.org/10.37745/ejcsit.2013/vol13n2121

Kamatchi, K., & Uma, E. (2025). Insights into user behavioral-based insider threat detection: systematic review. International Journal of Information Security, 24(2), 88.

Lu, Y., Fan, X., Zhang, Y., Wang, Y., & Jiang, X. (2023). Machine learning models using SHapley Additive exPlanation for fire risk assessment mode and effects analysis of stadiums. Sensors, 23(4), 2151.

Mienye, I. D., & Swart, T. G. (2024). A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications. Information, 15(12), 755. https://doi.org/10.3390/info15120755

Javed, H., El-Sappagh, S., & Abuhmed, T. (2024). Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artificial Intelligence Review, 58(1), 12. https://doi.org/10.1007/s10462-024-11005-9

Ofori, H. K., Bell-Dzide, K., Brown-Acquaye, W. L., Lempogo, F., Frimpong, S. O., Agbehadji, I. E., & Millham, R. C. (2025). Application of Machine Learning and Deep Learning Techniques for Enhanced Insider Threat Detection in Cybersecurity: Bibliometric Review. Symmetry, 17(10), 1704. https://doi.org/10.3390/sym17101704

Raftopoulos, G., Davrazos, G., & Kotsiantis, S. (2025). Evaluating Fairness Strategies in Educational Data Mining: A Comparative Study of Bias Mitigation Techniques. Electronics, 14(9), 1856. https://doi.org/10.3390/electronics14091856

Rathore, A. (2025). What Is an Insider Threat? Types, Examples, and Prevention. Retrieved January 10 from https://www.astrill.com/blog/what-is-an-insider-threat/#:~:text=1.,detect%20subtle%20misuse%20of%20privileges.

Reddy, C. K. K., Reddy, P. A., Reddy, P. S., Shuaib, M., Alam, S., Ahmad, S., & Rajaram, A. (2025). Twined ensemble framework for network security: integrating Random Forest, AdaBoost, and Gradient Boosting for enhanced intrusion detection. Discover Internet of Things, 5(1), 107.

Saminathan, K., Mulka, S. T. R., Damodharan, S., Maheswar, R., & Lorincz, J. (2023). An Artificial Neural Network Autoencoder for Insider Cyber Security Threat Detection. Future Internet, 15(12), 373. https://doi.org/10.3390/fi15120373

Sharma, A., Rani, S., & Shabaz, M. (2025). A comprehensive review of explainable AI in cybersecurity: Decoding the black box. ICT Express.https://doi.org/10.1016/j.icte.2025.10.004

Tao, X., Liu, J., Yu, Y., Zhang, H., & Huang, Y. (2025). An insider threat detection method based on improved Test-Time Training model. High-Confidence Computing. https://doi.org/10.1016/j.hcc.2024.100283

Wei, Y., Chow, K. P., & Yiu, S. M. (2020, January). Insider threat detection using multi-autoencoder filtering and unsupervised learning. In IFIP International Conference on Digital Forensics (pp. 273-290). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-56223-6_15

Yi, J., & Tian, Y. (2024). Insider Threat Detection Model Enhancement Using Hybrid Algorithms between Unsupervised and Supervised Learning. Electronics, 13(5), 973. https://doi.org/10.3390/electronics13050973

Zheng, P., Yuan, S., & Wu, X. (2021). Using Dirichlet marked Hawkes processes for insider threat detection. Digital Threats: Research and Practice (DTRAP), 3(1), 1-19. https://doi.org/10.1145/3457908