



Received: 27-08-2026
Accepted: 07-10-2026

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Weighted Percentile-Rank Fusion of Heterogeneous Unsupervised Anomaly Detection Models for Insider Threat Detection

¹ Musonda Dyson, ² Sesu Ovid

^{1,2} Department of Computer Science, DMI St. Eugene University, Chibombo, Zambia

Corresponding Author: **Musonda Dyson**

Abstract

Insider threat detection is challenging because malicious activity may be performed through legitimate user accounts and can therefore resemble normal organisational behaviour. Unsupervised anomaly detection identifies behavioural deviations without requiring malicious labels during training, however different detection algorithms may produce heterogeneous anomaly scores and identify different aspects of anomalous behaviour. This study investigates a weighted percentile-rank score-fusion approach for combining four heterogeneous unsupervised anomaly detectors within a common behavioural analytics framework. The study uses the CERT Insider Threat Dataset Release 4.2 and represents user activity at the user-week level using behavioural features derived from logon, device, file, web, and email activity. Isolation Forest, a stochastic gradient descent one-class support vector machine, a copula-based outlier detector, and MiniBatch K-Means clustering are independently trained and evaluated. Their anomaly

scores are transformed into empirical percentile ranks before weighted aggregation, with ensemble weights selected using validation data and subsequently fixed for held-out testing. On 27,302 test user-weeks containing 157 malicious user-weeks, the weighted ensemble achieved an area under the precision-recall curve of 0.094 and an area under the receiver operating curve of 0.899, exceeding the precision-recall area of every individual detector. Leave-one-detector-out analysis showed that excluding the copula-based detector reduced the precision-recall area to 0.072, whereas excluding the clustering-based detector increased it to 0.097. Weight-sensitivity experiments further examined individual detector contributions to ensemble behaviour. The findings demonstrate both the potential and limitations of heterogeneous score fusion for highly imbalanced insider-threat detection and provide empirical evidence on how detector composition affects the resulting ensemble.

Keywords: Insider Threat Detection, User and Entity Behaviour Analytics, Anomaly Detection, Unsupervised Machine Learning, Score-Level Fusion, Percentile-Rank Fusion

1. Introduction

Insider threats represent a distinctive cybersecurity problem because potentially harmful activity can be performed through legitimate user accounts and authorised access mechanisms [1, 2]. Unlike attacks that necessarily involve unauthorised access or recognisable malicious signatures, insider-related activity may consist of individually legitimate actions whose security significance becomes apparent only when their behavioural context is considered. File access, removable-device usage, email communication, web activity, and authentication events can all occur during normal organisational work. Consequently, detecting insider threats requires methods capable of identifying meaningful deviations from expected behavioural patterns rather than relying exclusively on predefined signatures or isolated security events.

User Behavior Analytics (UBA), and more broadly User and Entity Behavior Analytics (UEBA), provides a behavioural perspective for this problem [4]. Rather than examining individual events independently, behavioural analytics can aggregate activity into representations describing how users interact with organisational systems over time. Such representations can capture activity frequency, volume, temporal patterns, and changes relative to previous behaviour. Machine learning-based anomaly detection is particularly relevant because insider-threat labels are typically limited, making fully supervised approaches difficult to apply without relying heavily on assumptions about known malicious behaviour [5].

Unsupervised anomaly detection algorithms provide different mechanisms for identifying observations that deviate from learned behavioural patterns. An isolation-based detector, a clustering-based detector, a boundary-based detector, and a

statistical detector may therefore respond differently to the same behavioural population. In this study, four such approaches are considered. Isolation Forest [8], MiniBatch K-Means [17], SGD One-Class Support Vector Machine [18], and Copula-Based Outlier Detection (COPOD) [7]. The purpose of using these detectors together is not to assume that one algorithm is universally superior, but to investigate whether structurally different detection mechanisms identify complementary aspects of anomalous behaviour.

The use of multiple detectors introduces an important score-integration problem. The four algorithms generate anomaly outputs according to different scoring mechanisms, and their numerical scales are not necessarily directly comparable. A detector with numerically larger scores should not automatically receive greater influence in an ensemble merely because of the scale of its output. Direct arithmetic aggregation of heterogeneous raw scores can therefore produce a fusion result that is influenced by score magnitude rather than by the relative anomalousness represented by each detector [11].

Score normalisation provides a means of addressing this problem. In the present study, detector outputs are transformed into empirical percentile ranks before fusion, consistent with rank-based approaches identified in the outlier-ensemble literature [10, 11]. Each observation is consequently represented according to its relative position within the score distribution of the corresponding detector. The normalised scores can then be combined using explicitly defined weights. This approach separates two distinct operations: determining how anomalous an observation appears relative to each detector's scoring distribution, and determining how much each detector contributes to the final fused score.

The central question of this study is therefore not simply whether an ensemble produces a higher metric than an individual model. Instead, the study investigates whether heterogeneous detector outputs can be transformed into a common ranking representation and combined in a controlled manner, and whether the resulting fusion provides useful detection behaviour relative to the constituent algorithms. This distinction is important because ensemble methods do not inherently guarantee improved performance. A detector may contribute complementary information, redundant information, or information that reduces the performance of the combined system.

The study uses the CERT Insider Threat Dataset Release 4.2 as the experimental data source [3]. The dataset provides multiple forms of simulated organisational activity together with documented insider-threat scenarios and associated ground truth information. The availability of ground truth allows the anomaly detectors to remain unsupervised during model fitting while permitting their resulting rankings to be evaluated against known malicious activity. This separation between unsupervised learning and labelled evaluation is maintained throughout the experimental design.

The behavioural representation used in the study is constructed at user-week granularity from five activity modalities: logon, device, file, HTTP, and email. The same representation is provided to all four detectors so that differences in their results can be examined without simultaneously changing the underlying feature representation. The detector outputs are subsequently converted to percentile ranks and combined through a weighted fusion mechanism. The ensemble weights are

selected using the validation partition and frozen before the final test evaluation.

The experimental analysis extends beyond a single aggregate performance measure. Individual detectors and the fused ensemble are evaluated using receiver operating characteristic area under the curve (ROC-AUC) and precision-recall area under the curve (PR-AUC). Additional experiments examine the effect of removing individual detectors from the ensemble and the sensitivity of performance to changes in detector weights. Threshold-based detection, ranking at selected operating points, and persistence are also considered to provide a broader view of detection behaviour under the highly imbalanced test population.

A single unsupervised anomaly detector provides only one analytical view of user behavioural abnormality. Because anomaly detection algorithms rely on different structural assumptions, they may identify different subsets of observations as anomalous, so relying on one detector may fail to capture behavioural deviations that are more readily identified by another detection paradigm. Using multiple detectors does not, however, directly solve the problem on its own: the resulting anomaly scores may have substantially different numerical scales and distributions, and simple raw-score averaging can therefore give disproportionate influence to a detector because of its scoring scale rather than because of meaningful detector agreement. There is consequently a need for a score-fusion mechanism that first establishes a comparable representation of anomaly extremeness and then combines the resulting detector outputs according to explicitly defined contributions. The problem addressed by this study is therefore whether heterogeneous unsupervised anomaly detector outputs can be transformed through percentile ranking and combined using a validation-selected weighted fusion mechanism to provide useful insider-threat detection performance under a common behavioural representation.

This problem reflects three related gaps in the existing literature. First, different unsupervised anomaly detection paradigms can encode different notions of abnormality, yet their comparative behaviour can be difficult to establish when studies use different datasets, feature representations, or experimental protocols, applying structurally different detectors to the same behavioural representation provides a controlled basis for comparison. Second, multiple detectors can produce anomaly scores with different numerical characteristics, and combining such scores requires an explicit normalisation strategy if detector influence is not to depend primarily on arbitrary differences in score scale, for which percentile transformation provides a rank-based representation in which scores express relative extremeness within their respective detector distributions [11]. Third, the use of an ensemble should not itself be interpreted as evidence of improved detection [10]. A fused model may outperform individual detectors, reproduce their behaviour, or perform worse depending on detector complementarity and the selected weighting scheme [6, 11], so empirical comparison, leave-one-detector-out ablation, and weight-sensitivity analysis are required to determine how the constituent detectors affect the fused result. This study addresses these gaps by applying four heterogeneous unsupervised anomaly detection algorithms to a common user-week behavioural representation, transforming their outputs into empirical percentile ranks, and evaluating a

validation-selected weighted fusion against its constituent detectors.

The aim of this study is accordingly to investigate and evaluate a weighted percentile-rank score-fusion approach for combining heterogeneous unsupervised anomaly detection models for insider-threat detection within a common user behaviour analytics framework. To achieve this aim, the study evaluates Isolation Forest, SGD One-Class Support Vector Machine, COPOD, and MiniBatch K-Means using a common user-week behavioural representation derived from the CERT Insider Threat Dataset Release 4.2, transforms the heterogeneous anomaly scores produced by the four detectors into empirical percentile-rank representations suitable for score-level fusion, constructs a weighted ensemble of the normalised detector outputs using weights selected from validation data and subsequently fixed for held-out evaluation, compares the detection performance of the weighted ensemble with that of the individual anomaly detection algorithms, examines detector contribution through leave-one-detector-out ablation and sensitivity to changes in ensemble weights, and characterises the detection behaviour of the individual detectors and ensemble using ranking-based and threshold-oriented evaluation measures.

Within this aim, the study addresses five research questions. RQ1 asks how effectively Isolation Forest, MiniBatch K-Means, SGD One-Class Support Vector Machine, and COPOD detect insider-threat-related behavioural anomalies under a common user-week representation. RQ2 asks how weighted percentile-rank score fusion compares with the individual anomaly detectors. RQ3 asks whether percentile-rank transformation provides a suitable basis for combining heterogeneous anomaly scores compared with alternative score-normalisation and fusion approaches. RQ4 asks to what extent the contribution of individual detectors affects the performance of the weighted ensemble. RQ5 asks how different operating configurations affect the precision, recall, alert volume, and user-level coverage of the weighted ensemble.

The primary contribution of this study is an empirical investigation of weighted percentile-rank fusion for heterogeneous unsupervised anomaly detection in insider-threat behavioural analytics. The study provides a common multimodal behavioural representation through which four structurally different anomaly detection paradigms can be compared under the same experimental conditions, together with a rank-based score transformation that places heterogeneous detector outputs into a common relative scale before weighted aggregation. A second contribution is the controlled evaluation of the resulting fusion rather than treating ensemble construction as inherently beneficial. Individual-detector comparison, leave-one-detector-out ablation, and weight-sensitivity analysis are used to examine how the composition of the ensemble affects detection performance, providing evidence not only about the performance of the fused score but also about the contribution and interaction of its constituent detectors. Finally, the study evaluates the fusion approach using both ranking-based and threshold-oriented measures, which is particularly important for insider-threat detection because the positive class is substantially smaller than the normal behavioural population, the resulting analysis therefore considers the distinction between ranking anomalous

observations and achieving a useful operating point for detection.

2. Materials and Methods

2.1 Research Design

This study adopted a quantitative experimental research design to investigate the behaviour of heterogeneous unsupervised anomaly detection algorithms and the effectiveness of combining their outputs through weighted percentile-rank fusion. The experimental procedure was applied to a fixed behavioural dataset and evaluated against documented insider-threat ground truth.

The methodology consisted of five sequential stages: dataset preparation and preprocessing, behavioural feature engineering, independent anomaly detection, score transformation and ensemble fusion, and empirical evaluation. A common behavioural representation was constructed once and supplied to all four detectors. This controlled design allowed differences in detector performance and ensemble behaviour to be examined without simultaneously changing the underlying feature representation.

The detection methodology was unsupervised. The four constituent algorithms were fitted without supplying malicious or benign labels as training targets. The CERT ground-truth answer key was maintained separately and was used for evaluation and validation-stage methodological selection. Thus, the study represents an unsupervised anomaly-detection methodology evaluated against labelled ground truth.

2.2 Dataset

The study used the CMU CERT Insider Threat Dataset Release 4.2^[3]. The dataset simulates digital activity within a fictitious organisation and contains approximately 1,000 users over an observation period extending from 1 December 2009 to 31 May 2011. Five principal behavioural modalities were used: logon and logoff activity, removable-device activity, file activity, HTTP activity, and email activity.

The dataset contains documented simulated insider-threat scenarios and an associated answer key identifying malicious users and their corresponding activity windows. The answer key was not supplied to the anomaly detection algorithms during model fitting. Instead, it provided the reference labels required to evaluate the detector rankings and to perform validation-stage selection of the ensemble configuration.

The final behavioural feature artifact contained 67,238 user-week observations and 88 columns. Of these, 86 were modelling features, with 'user_id' and 'week_start' retained as identifying fields. The modelling population represented 1,000 unique users. The final prediction population used for evaluation contained 32,920 user-week observations.

The unit of analysis was therefore the user-week rather than the individual event. This aggregation provides a behavioural context in which activity from multiple telemetry sources can be represented jointly while retaining a temporal modelling structure.

2.3 Data preprocessing

Before feature engineering, records from the five behavioural modalities were parsed, cleaned, and

consolidated into a unified activity representation. Source-specific fields were standardised to provide consistent identifiers and timestamps across the activity streams. Each event retained its originating behavioural modality so that modality-specific activity could be incorporated into the subsequent feature representation.

The raw data contained substantial differences in event frequency between modalities. HTTP activity, for example, occurred at a substantially higher volume than logon or device activity. Consequently, the methodology did not treat the raw event counts themselves as directly comparable behavioural variables. Instead, the activity streams were transformed into aggregated behavioural measures appropriate to the user-week unit of analysis.

Ground-truth labels were constructed separately from the CERT answer key. This separation was important because the objective was to evaluate unsupervised anomaly detection rather than supervised classification. The labels were therefore excluded from the feature matrix and from detector fitting.

2.4 Behavioural feature engineering

The consolidated activity data were transformed into an 86-feature behavioural representation at user-week granularity. The feature representation comprised three principal groups. base behavioural features, temporal and multi-timescale features, and week-on-week features.

The base behavioural features represented absolute characteristics of user activity during the modelling period. These included measures derived from authentication, device, file, HTTP, and email activity. Examples included activity counts, after-hours activity, weekend activity, unique computer usage, removable-device activity, file activity, web activity, email activity, attachment-related activity, and external-recipient behaviour.

Temporal and multi-timescale features were designed to capture when activity occurred and how it was distributed within the modelling period. These features incorporated behavioural timing and concentration characteristics that could distinguish a user's activity pattern from another user's pattern even when their overall activity volumes were similar.

The third group comprised week-on-week features. These features represented changes in behavioural activity relative to preceding observations and were intended to capture deviations from an individual's recent behavioural baseline. The combination of absolute, temporal, and change-oriented features allowed the representation to encode both the level of activity and changes in behaviour over time.

The resulting representation was held constant across the four anomaly detection algorithms. This was a central control in the experimental design because it ensured that observed differences between algorithms were attributable primarily to their detection mechanisms rather than to differences in feature construction.

2.5 Feature preprocessing and scaling

Feature preprocessing was performed before anomaly detection. Missing numerical feature values were handled using median imputation, with the imputation parameters derived from the training partition. The resulting features were standardised using a StandardScaler, implemented via the scikit-learn machine-learning library ^[16], fitted exclusively on the training data.

The use of training-fitted preprocessing prevented information from the validation and test partitions from influencing the transformation parameters used during model training. The same fitted preprocessing transformation was subsequently applied to the validation and test observations.

It is important to distinguish this feature-level scaling from the later anomaly-score normalisation. Standardisation was performed to place model input variables on an appropriate numerical scale. Percentile-rank transformation, by contrast, was applied to the outputs of the trained anomaly detectors so that heterogeneous anomaly scores could be combined.

2.6 Chronological data partitioning

A chronological three-way partition was used to separate model fitting, methodological selection, and final evaluation. The complete feature population covered the period from 28 December 2009 to 16 May 2011 at the user-week representation.

The training partition contained observations through 31 August 2010. The validation partition covered observations after 31 August 2010 and through 15 October 2010. The test partition contained observations after 15 October 2010.

Table A: Chronological partitioning of the behavioural feature population

Partition	User-weeks	Purpose
Training	34,318	Model fitting and preprocessing
Validation	5,618	Ensemble selection and threshold selection
Test	27,302	Final held-out evaluation
Total	67,238	Complete feature population

The chronological ordering was maintained throughout the experiment. In particular, test observations were not used to select detector weights, normalisation references, or operating thresholds.

2.7 Individual anomaly detection algorithms

Four unsupervised anomaly detection algorithms representing different detection paradigms were evaluated. The detectors were trained independently using the same preprocessed behavioural representation.

2.7.1 Isolation Forest

Isolation Forest ^[8] is an isolation-based anomaly detection method that identifies observations according to how readily they can be separated from other observations through recursive random partitioning. Observations that are isolated using relatively short partition paths are considered more anomalous.

Within this study, Isolation Forest served as the primary reference detector against which the other approaches and the fused ensemble were examined. Its output was retained as an independent detector score before participating in the percentile-rank fusion process.

2.7.2 MiniBatch K-Means

MiniBatch K-Means ^[17] represents a clustering-based approach to anomaly detection. The algorithm learns cluster structures within the behavioural feature space using mini-batches, allowing clustering to be performed more efficiently on larger datasets. Observations that are relatively distant from their representative cluster structure can consequently be treated as more anomalous.

Its inclusion provided a clustering perspective that differed from the isolation, boundary, and distribution-based

approaches represented by the other detectors.

2.7.3 SGD One-Class Support Vector Machine

SGD One-Class Support Vector Machine ^[18] provides a boundary-based approach in which a representation of expected observations is learned and observations outside the learned behavioural region can be identified as anomalous. The stochastic-gradient implementation allows the method to be applied efficiently to the resulting feature population.

The detector provided a complementary boundary-oriented perspective on behavioural abnormality.

2.7.4 Copula-Based Outlier Detection

Copula-Based Outlier Detection (COPOD) ^[7] is a statistical anomaly detection method that evaluates observations using empirical distributional characteristics and tail probabilities. Rather than relying on isolation or explicit clustering, COPOD assesses how unusual observations are in relation to the empirical distributions of the input variables.

The inclusion of COPOD provided a distribution-based perspective that was structurally different from the other three detectors.

2.8 Independent detector scoring

Each detector was fitted independently on the training behavioural representation. The resulting models generated anomaly scores for the validation and test observations.

The four detectors do not inherently produce scores on an identical numerical scale. Their score distributions reflect the internal scoring mechanisms of their respective algorithms. Consequently, the raw outputs were retained for individual-detector evaluation but were not directly averaged to produce the final ensemble.

This distinction is important because numerical magnitude alone does not establish that a score from one detector represents a greater degree of anomalousness than an identically sized score from another detector.

2.9 Percentile-rank score transformation

The principal methodological component of the study was the transformation of heterogeneous anomaly scores into a common percentile-rank representation before fusion.

Let $s_i(x)$ denote the anomaly score assigned to observation x by detector i . Rather than combining $s_i(x)$ directly with scores produced by other detectors, the score was represented according to its empirical relative position within the corresponding detector score distribution.

The percentile representation can be expressed generally as.

$$p_i(x) = F_i(s_i(x))$$

Where F_i represents the empirical cumulative distribution function associated with detector i 's reference scores. Equivalently, the transformation can be understood as expressing an observation's score as a percentile within the detector-specific reference distribution.

The resulting percentile values place detector outputs on a common relative scale. A value near the upper end of the percentile range therefore indicates that an observation is highly extreme relative to the reference distribution for that detector, regardless of the detector's original numerical score scale.

The percentile transformation was based on validation-period reference distributions and was subsequently kept fixed for the held-out test evaluation. This prevented the test

population from determining the normalisation reference used to evaluate itself.

The approach therefore separates three operations. (1) generation of an anomaly score by each independent detector, (2) transformation of each score into a comparable percentile representation, and (3) weighted aggregation of the normalised representations.

This score-level architecture differs from an end-to-end ensemble classifier because the constituent detectors remain independently trained and the fusion occurs only after their anomaly scores have been generated.

2.10 Weighted percentile-rank fusion

Following percentile transformation, the four detector representations were combined using a weighted sum.

For an observation x , the ensemble score was defined as.

$$S_{ensemble}(x) = \sum_{i=1 \text{ to } 4} w_i \cdot p_i(x)$$

Subject to.

$$\sum_{i=1 \text{ to } 4} w_i = 1, w_i \geq 0$$

Table B: Frozen ensemble weights selected using the validation partition

Detector	Ensemble weight
Isolation Forest	0.10
COPOD	0.50
MiniBatch K-Means	0.30
SGD One-Class SVM	0.10
Total	1.00

The weights were selected using the validation partition and subsequently frozen before the final test evaluation. The resulting ensemble therefore did not use test performance to determine the relative contribution of the four detectors.

The selected configuration assigns the largest contribution to COPOD, followed by MiniBatch K-Means, while Isolation Forest and SGD One-Class SVM each contribute 0.10. These values are treated as an empirical configuration produced by the validation procedure rather than as assumptions about the general superiority of any detector.

2.11 Experimental evaluation

The evaluation was designed to assess both individual detector performance and the behaviour of the fused ensemble.

The primary ranking measures were Average Precision, reported as precision-recall area under the curve (PR-AUC), and receiver operating characteristic area under the curve (ROC-AUC). PR-AUC was particularly important because the test population was highly imbalanced, containing 157 malicious user-weeks among 27,302 evaluated observations, PR-AUC is widely recommended over ROC-AUC alone as the primary ranking measure under such imbalance ^[15].

Threshold-based evaluation was also performed using thresholds selected from validation data and subsequently frozen for test evaluation. Precision, recall, F1-score, confusion-matrix components, alert volume, and user-level coverage were examined at the selected operating point.

Additional ranking-oriented analyses included Precision@10, Precision@25, Precision@50, and Precision@100. A Top-1% operating analysis was also conducted to examine the behaviour of the highest-ranked

observations without relying exclusively on a fixed threshold.

A two-week persistence analysis was used to examine whether requiring repeated anomalous observations altered the precision-recall and alert-volume characteristics of the detectors and ensemble.

2.12 Leave-one-detector-out ablation

A leave-one-detector-out ablation experiment was conducted to examine the contribution of individual detectors to the fused ensemble.

Five configurations were evaluated. (1) the complete four-detector ensemble, (2) the ensemble without Isolation Forest, (3) the ensemble without SGD One-Class SVM, (4) the ensemble without COPOD, and (5) the ensemble without MiniBatch K-Means.

When a detector was removed, the remaining weights were renormalised so that the total ensemble weight remained equal to one. The experiment therefore examined how the fused result changed when each constituent detector was excluded while retaining a valid weighted combination of the remaining detectors.

This experiment was used to investigate detector complementarity rather than to establish a simple ranking of algorithms. A change in performance following detector removal provides evidence about the information contributed by that detector within the particular ensemble configuration.

2.13 Weight-sensitivity analysis

A weight-sensitivity experiment was conducted to examine how changes in the contribution assigned to individual detectors affected ensemble performance.

The analysis varied the weight of one detector around the authoritative configuration while rescaling the remaining detector weights. The evaluated configurations included detector weights ranging from 0.0 to 0.7 across the four detector-specific sensitivity analyses.

The purpose of this experiment was to determine whether the selected ensemble configuration represented a highly sensitive operating point or whether similar performance could be obtained across a broader range of detector contributions. Test PR-AUC and ROC-AUC, together with validation PR-AUC, were retained as the principal measures for interpreting the sensitivity results.

2.14 Leakage control and reproducibility

Several controls were incorporated to preserve the separation between model development and final evaluation. First, the four anomaly detectors were fitted without ground-truth labels. Second, preprocessing parameters were fitted using the training partition. Third, percentile reference distributions and ensemble weights were determined using validation data rather than the test population. Fourth, the ensemble configuration and threshold were frozen before final test evaluation. Finally, the test observations were used only to obtain the reported final performance measurements. The experimental implementation used deterministic random-state configuration for the stochastic machine learning procedures. The principal project configuration specified a random state of 42, while the model and experiment infrastructure maintained explicit random-seed configuration for reproducibility.

These controls are important because the purpose of the

final test evaluation was to estimate the behaviour of a configuration that had already been determined without access to test outcomes. The methodology therefore distinguishes methodological selection from final performance assessment.

2.15 Score-normalisation comparison

To examine whether the percentile-rank transformation provided a suitable basis for heterogeneous score fusion relative to alternative normalisation strategies, two additional fusion configurations were constructed using the same four detectors and the same frozen detector weights (0.10 Isolation Forest, 0.50 COPOD, 0.30 MiniBatch K-Means, 0.10 SGD One-Class SVM) described in Section 2.10, but with the percentile-rank transformation replaced by an alternative score-normalisation step. This design held the detector weighting constant across all three configurations so that the comparison isolated the effect of the normalisation strategy rather than conflating it with a change in detector weighting.

The first alternative configuration, raw-score weighted fusion, applied the frozen weights directly to each detector's untransformed anomaly score, without any normalisation step. The second alternative, validation-fitted Min-Max weighted fusion, rescaled each detector's score to the zero-to-one range using minimum and maximum values obtained from the validation partition, analogous to the validation-period reference distribution used for the percentile-rank transformation, before applying the frozen weights. Both alternative configurations were evaluated identically to the percentile-rank ensemble, using validation PR-AUC together with test PR-AUC and test ROC-AUC.

2.16 Evaluation population and ground truth

The complete prediction artifact contained 32,920 user-week observations. Of these, 5,618 belonged to the validation partition and 27,302 belonged to the test partition.

The authoritative ground-truth labels identified 213 malicious user-week observations within the prediction population. Of these, 56 occurred in the validation partition and 157 occurred in the test partition. The remaining observations in the evaluation population were treated as non-malicious for the purpose of the binary performance calculations.

The test population therefore consisted of 27,302 observations, including 157 malicious and 27,145 non-malicious user-weeks. This substantial class imbalance was explicitly considered when interpreting the resulting detection metrics.

The distinction between user-level and user-week-level labels was also retained. A malicious user may have multiple normal user-weeks outside the documented malicious activity window. Consequently, the evaluation measures describe the ability to identify malicious user-week observations, rather than implying that every observation generated by a maliciously assigned user should be anomalous.

3. Results and Discussion

3.1 Evaluation population and dataset characteristics

The behavioural feature engineering process produced 67,238 user-week observations across 1,000 users, represented by 86 modelling features together with the identifying fields 'user_id' and 'week_start'. The

chronological partitioning procedure allocated 34,318 user-weeks (2009-12-28 to 2010-08-30) to the training partition, 5,618 user-weeks (2010-09-06 to 2010-10-11) to the validation partition, and 27,302 user-weeks (2010-10-18 to 2011-05-16) to the test partition.

The evaluation population, comprising the validation and test partitions used to assess detector and ensemble performance, totalled 32,920 user-weeks. Within this population, the CERT answer key identified 213 malicious user-week observations. 56 in the validation partition and 157 in the test partition. The test partition therefore contained 157 malicious user-weeks among 27,302 observations, an imbalance ratio of approximately 1 malicious user-week per 173 non-malicious user-weeks (a malicious-class prevalence of approximately 0.575 percent). This imbalance is consistent with the operational reality that insider-threat activity constitutes a small minority of overall user behaviour and motivated the selection of PR-AUC, rather than ROC-AUC alone, as the primary ranking measure throughout the evaluation ^[15] (RQ1).

3.2 Individual detector performance

Table 1 reports the test-partition PR-AUC and ROC-AUC achieved by each of the four independently trained anomaly detection algorithms.

Table 1: Individual detector performance on the test partition (27,302 user-weeks, 157 malicious)

Detector	PR-AUC	ROC-AUC
Isolation Forest	0.080834	0.870981
SGD One-Class SVM	0.074971	0.838922
COPOD	0.091153	0.936180
MiniBatch K-Means	0.027260	0.810095

All four detectors achieved ROC-AUC values substantially above the 0.50 level expected of random ranking, indicating that each algorithm captured some ranking signal relevant to the malicious user-weeks despite being fitted without access to ground-truth labels. COPOD produced the highest individual PR-AUC (0.091153) and ROC-AUC (0.936180) of the four detectors, followed by Isolation Forest. MiniBatch K-Means produced the lowest PR-AUC (0.027260) and the lowest ROC-AUC (0.810095) among the four algorithms, suggesting that the clustering-based representation of normality was comparatively less discriminative of the malicious user-weeks within this behavioural feature space than the statistical, isolation-based, and boundary-based perspectives.

These results should be interpreted as evidence about detector behaviour under the specific 86-feature user-week representation and chronological partitioning used in this study, rather than as a general ranking of the four algorithmic paradigms. The comparatively strong performance of COPOD indicates that the malicious user-weeks in this dataset were frequently associated with statistically extreme values along one or more behavioural dimensions, a characteristic that a distribution-based detector is well positioned to identify.

3.3 Detector score correlation

To examine the extent to which the four detectors' anomaly scores carried independent or overlapping ranking information, Spearman rank correlation coefficients were computed between every pair of detectors on the test

partition, using both the untransformed raw anomaly scores and the percentile-rank-transformed scores subsequently used in the fusion step described in Section 2.10. Fig 1 presents the percentile-rank correlation matrix as a heatmap. The corresponding raw-score correlation matrix was numerically almost identical to the percentile-rank matrix (differing only from the sixth decimal place onward), which is expected because the percentile-rank transformation is a monotonic function of each detector's raw score and therefore preserves Spearman rank correlation by construction.

Isolation Forest and MiniBatch K-Means formed the most strongly correlated pair of detectors (Spearman's $\rho = 0.945$), followed by Isolation Forest and COPOD ($\rho = 0.929$) and COPOD and MiniBatch K-Means ($\rho = 0.832$). SGD One-Class SVM was comparatively weakly correlated with each of the other three detectors ($\rho = 0.579$ - 0.629), indicating that its ranking of user-weeks diverged the most from the remaining detectors despite its intermediate individual PR-AUC reported above.

These correlation results provide a mechanistic explanation for the leave-one-detector-out ablation findings reported in Section 3.6. MiniBatch K-Means, the weakest individual detector by both PR-AUC and ROC-AUC, was also the detector most strongly correlated with two of the three remaining detectors, Isolation Forest and COPOD. Its removal from the ensemble therefore discarded a largely redundant, comparatively low-quality ranking signal rather than an independent one, which is consistent with the ablation finding that excluding MiniBatch K-Means improved rather than degraded both PR-AUC and ROC-AUC (Section 3.6). SGD One-Class SVM, by contrast, exhibited the lowest correlation with the other three detectors, indicating that it contributed the most independent ranking signal among the four detectors, notwithstanding its modest individual performance, this is consistent with its comparatively small but measurable ablation effect on PR-AUC (Section 3.6), since a weak but independent contributor can still provide a complementary signal that a stronger but more redundant detector does not. These findings reinforce the position, introduced in Section 2.11 and revisited in Section 3.6, that diversity of algorithmic paradigm does not by itself guarantee diversity of ranking behaviour, and that empirical correlation and ablation analysis together are necessary to establish which detectors provide genuinely complementary information within a fused ensemble.

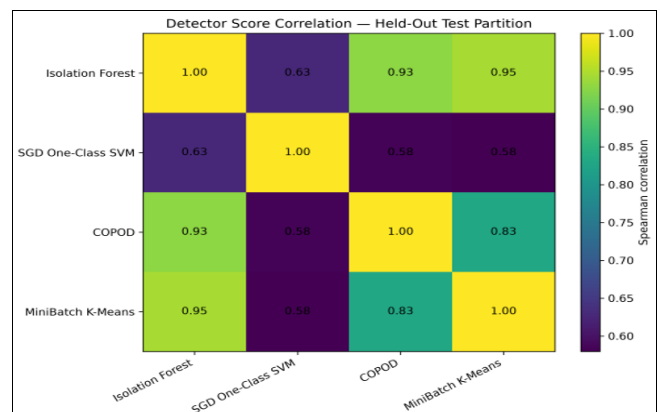


Fig 1: Spearman rank correlation between detector percentile-rank anomaly scores on the test partition

3.4 Weighted percentile-rank ensemble performance

The weighted percentile-rank ensemble, using the validation-selected and subsequently frozen weights of 0.10 (Isolation Forest), 0.50 (COPOD), 0.30 (MiniBatch K-Means), and 0.10 (SGD One-Class SVM), achieved a test PR-AUC of 0.094211 and a test ROC-AUC of 0.899354. Table 2 situates the ensemble result alongside the four individual detectors.

Table 2: Ensemble performance compared with individual detectors (test partition)

Model	PR-AUC	ROC-AUC
Isolation Forest	0.080834	0.870981
SGD One-Class SVM	0.074971	0.838922
COPOD	0.091153	0.936180
MiniBatch K-Means	0.027260	0.810095
Weighted percentile-rank ensemble	0.094211	0.899354

The ensemble PR-AUC of 0.094211 exceeded the PR-AUC of every individual detector, including COPOD (0.091153), the strongest individual contributor. The ensemble ROC-AUC of 0.899354, however, fell below the ROC-AUC achieved by COPOD alone (0.936180), although it exceeded the ROC-AUC of the remaining three detectors. This pattern indicates that the weighted percentile-rank fusion improved the ranking of malicious user-weeks specifically within the higher-precision region of the ranking that PR-AUC emphasises, while not uniformly improving ranking performance across the full operating range captured by ROC-AUC (RQ2, RQ3).

This outcome is consistent with the methodological position adopted in Section 2.11, namely that ensemble construction should not be assumed to produce uniform improvement across every evaluation measure ^[10]. The percentile-rank transformation placed the four heterogeneous detector outputs on a common relative scale prior to aggregation, and the resulting fused score provided the best PR-AUC observed among the evaluated configurations. Section 3.5 reports a direct comparison against alternative score-normalisation strategies that further addresses RQ3. The ROC-AUC result, by contrast, indicates that COPOD's own ranking of the full evaluation population was difficult for the fused configuration to fully reproduce once the remaining three, comparatively weaker, detectors were incorporated into the weighted combination.

Fig 2 presents the full precision-recall curves underlying the summary PR-AUC values reported in Table 2, computed directly from the stored per-user-week scores rather than from the summary statistic alone. The curves show that the weighted percentile-rank ensemble and COPOD traced broadly similar precision at low recall, with the ensemble curve remaining at or above every individual detector's curve across most of the recall range, consistent with the higher PR-AUC values reported above. All five curves converged toward the malicious-class prevalence of 0.0058 at full recall, reflecting the severe class imbalance of the evaluation population described in Section 3.1.

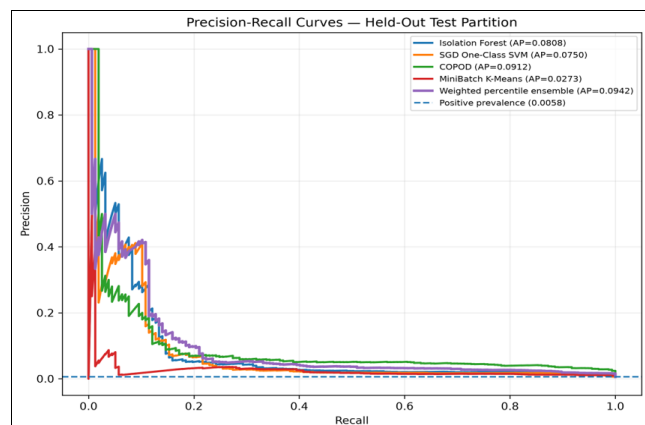


Fig 2: Precision-recall curves for the four individual detectors and the weighted percentile-rank ensemble on the test partition

3.5 Score-normalisation comparison

To test whether the percentile-rank transformation itself was responsible for the ensemble's performance, rather than the choice of detector weights, two alternative fusion configurations were evaluated using the identical frozen weights (0.10 Isolation Forest, 0.50 COPOD, 0.30 MiniBatch K-Means, 0.10 SGD One-Class SVM) described in Section 2.10. raw-score weighted fusion, in which the frozen weights were applied directly to each detector's untransformed anomaly score, and validation-fitted Min-Max weighted fusion, in which each detector's score was rescaled to the zero-to-one range using validation-partition minimum and maximum values before the frozen weights were applied. Table 3 reports the resulting validation PR-AUC, test PR-AUC, and test ROC-AUC for all three fusion strategies.

Table 3: Score-normalisation and fusion method comparison, using identical frozen detector weights

Fusion method	Validation PR-AUC	Test PR-AUC	Test ROC-AUC
Raw-score weighted fusion	0.218564	0.087859	0.928916
Validation-fitted Min-Max weighted fusion	0.219017	0.082397	0.911138
Weighted percentile-rank fusion	0.228673	0.094211	0.899354

Percentile-rank fusion achieved the highest PR-AUC of the three normalisation strategies on both the validation partition (0.228673, compared with 0.219017 for Min-Max and 0.218564 for raw-score fusion) and the test partition (0.094211, compared with 0.082397 for Min-Max and 0.087859 for raw-score fusion), a relative test PR-AUC improvement of approximately 14.3 percent over Min-Max fusion and 7.2 percent over raw-score fusion. This directly addresses RQ3. under identical detector weighting, replacing the percentile-rank transformation with either raw-score or Min-Max normalisation reduced PR-AUC on both validation and test data, indicating that the percentile-rank transformation itself, rather than the frozen weight

configuration alone, contributed to the ensemble's precision-recall ranking performance.

The pattern was reversed for test ROC-AUC, where percentile-rank fusion produced the lowest value of the three strategies (0.899354), compared with 0.911138 for Min-Max fusion and 0.928916 for raw-score fusion. This indicates that percentile-rank transformation provided a specific advantage in the higher-precision region of the ranking that PR-AUC emphasises, rather than a uniform advantage across the complete ranking captured by ROC-AUC. Consequently, the percentile-rank transformation is best described as the normalisation strategy that produced the strongest precision-recall ranking performance among those evaluated, rather than as uniformly superior to raw-score or Min-Max fusion on every measure.

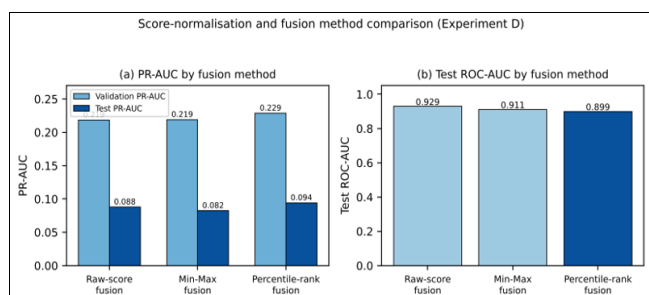


Fig 3: Score-normalisation and fusion method comparison under identical frozen detector weights (Experiment D)

3.6 Leave-one-detector-out ablation

Table 4 reports the leave-one-detector-out ablation results, in which each detector was removed from the ensemble in turn and the remaining weights were renormalised to sum to one.

Table 4: Leave-one-detector-out ablation (test partition, weights renormalised after removal)

Configuration	PR-AUC	ROC-AUC
Full ensemble (all four detectors)	0.094211	0.899354
Without Isolation Forest	0.094118	0.901472
Without SGD One-Class SVM	0.090245	0.899780
Without COPOD	0.071852	0.845594
Without MiniBatch K-Means	0.097079	0.923315

Removing COPOD produced the largest reduction in PR-AUC, from 0.094211 to 0.071852, a decrease of approximately 23.7 percent, and the largest reduction in ROC-AUC, from 0.899354 to 0.845594. This result confirms that COPOD was the dominant contributor to the fused ensemble's detection performance, consistent with its position as the highest-weighted detector (0.50) and its strongest individual-detector results reported in Section 3.2 (RQ4).

Removing SGD One-Class SVM produced a smaller but still measurable reduction in PR-AUC, from 0.094211 to 0.090245, while removing Isolation Forest left PR-AUC essentially unchanged (0.094211 to 0.094118) and marginally improved ROC-AUC (0.899354 to 0.901472). Most notably, removing MiniBatch K-Means increased both PR-AUC, from 0.094211 to 0.097079, and ROC-AUC, from 0.899354 to 0.923315, indicating that MiniBatch K-Means, the weakest individual detector reported in Section 3.2, reduced rather than improved the performance of the four-detector ensemble under the frozen weight configuration.

This finding is an important empirical result rather than a negative one. It demonstrates that the leave-one-detector-out procedure, applied after the weights had already been frozen using validation data, is capable of identifying detectors whose inclusion does not benefit the fused result under the selected configuration. It does not by itself imply that a three-detector ensemble excluding MiniBatch K-Means would necessarily be superior under an independently reselected weighting, because the ablation weights were obtained by renormalising the existing frozen weights rather than by repeating the validation-based weight-selection procedure for each reduced detector subset. The result nevertheless provides direct evidence addressing RQ4, showing that the four constituent detectors contributed unequally, and in the case of MiniBatch K-Means, counter-productively, to the frozen ensemble configuration. This is consistent with the broader outlier-ensemble literature, which cautions that algorithmic diversity does not by itself guarantee complementary contribution [10, 11].

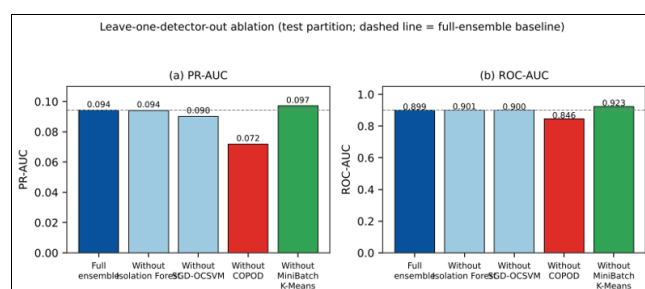


Fig 4: Leave-one-detector-out ablation on the test partition (dashed line marks the full-ensemble baseline)

3.7 Weight-sensitivity analysis

The weight-sensitivity experiment varied the contribution of each detector individually, across ranges of 0.00-0.30 for Isolation Forest, 0.30-0.70 for COPOD, 0.10-0.50 for MiniBatch K-Means, and 0.00-0.30 for SGD One-Class SVM, while rescaling the remaining three weights proportionally. The authoritative frozen configuration (0.10/0.50/0.30/0.10) was confirmed to be present within each sensitivity sweep, reproducing the reported test PR-AUC of 0.094211 and ROC-AUC of 0.899354 at that specific point. Fig 5 plots validation and test PR-AUC against the swept weight for each of the four detectors.

The four sensitivity curves exhibited distinct patterns. Increasing the COPOD weight from 0.30 to 0.70 produced a monotonic increase in both validation PR-AUC (0.219578 to 0.230582) and test PR-AUC (0.088313 to 0.097503), with no evidence of the curve turning downward within the explored range, indicating that the frozen weight of 0.50 was conservative relative to the direction of the trend rather than an overfitted peak. The Isolation Forest and SGD One-Class SVM sweeps instead showed an early local maximum close to the frozen configuration followed by decline. test PR-AUC for Isolation Forest peaked at a weight of 0.05 (0.094992) before falling to 0.091723 at a weight of 0.30, and test PR-AUC for SGD One-Class SVM similarly peaked at a weight of 0.05 (0.093414) before falling to 0.090923 at a weight of 0.30, consistent with these two detectors contributing a comparatively narrow band of useful weight around their frozen values of 0.10.

The MiniBatch K-Means sweep showed the clearest divergence between the validation and test partitions.

validation PR-AUC increased only marginally across the explored range (0.220413 at a weight of 0.10 to 0.229427 at a weight of 0.50), while test PR-AUC declined monotonically over the same range (0.096155 to 0.086529). This divergence provides additional, independent support for the ablation finding reported in Section 3.6 and the correlation-based explanation offered in Section 3.3, increasing MiniBatch K-Means' weight appeared mildly beneficial on the validation partition on which the weights were originally selected, but was consistently detrimental on the held-out test partition, reinforcing the conclusion that its frozen weight of 0.30 did not transfer the apparent validation-partition benefit to unseen data (RQ4).

The sensitivity ranges examined were constructed around the frozen configuration rather than spanning the full simplex of possible weight combinations, and they confirm that the selected weights do not sit at an extreme or arbitrary boundary of the explored space. In combination with the ablation results in Section 3.6, the sensitivity analysis supports the conclusion that COPOD's weight of 0.50 reflects its substantive contribution to ensemble performance, while the comparatively wide range explored for MiniBatch K-Means (0.10-0.50) is consistent with the finding that its presence in the ensemble, at the frozen weight of 0.30, did not provide a performance benefit relative to its exclusion (RQ4).

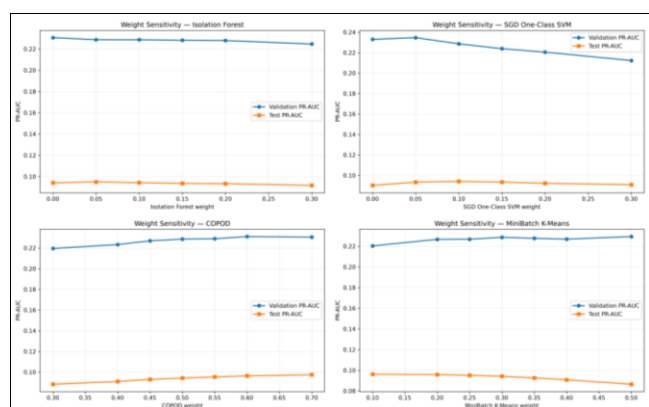


Fig 5: Weight-sensitivity analysis showing validation and test PR-AUC against the swept weight for (a) Isolation Forest, (b) SGD One-Class SVM, (c) COPOD, and (d) MiniBatch K-Means, with the remaining three weights rescaled proportionally

3.8 Threshold-based detection performance

A frozen operating threshold of 0.993402 was selected using the validation partition and subsequently applied without modification to the test partition. Table 5 reports the resulting confusion-matrix components and derived metrics at both partitions.

Table 5: Threshold-based detection performance at the frozen scientific threshold (0.993402)

Partition	Rows	Alerts	Alert rate (%)	Precision	Recall	F1-score
Validation	5,618	19	0.338199	0.473684	0.160714	0.240000
Test	27,302	82	0.300344	0.219512	0.114650	0.150628

On the test partition, the frozen threshold produced 82 alerts from 27,302 user-weeks (an alert rate of 0.300344 percent), with 18 true positives, 64 false positives, 139 false negatives, and 27,081 true negatives. Precision declined from 0.473684 on the validation partition to 0.219512 on the

test partition, while recall declined from 0.160714 to 0.114650. This reduction is consistent with a threshold selected under a relatively small validation sample (56 malicious user-weeks) being applied to a larger and temporally later test sample (157 malicious user-weeks), and illustrates the operational difficulty of translating ranking performance into a stable fixed-threshold alerting rule under severe class imbalance.

At the user level, the 82 test-partition alerts corresponded to 56 unique alerted users, of whom 14 were among the 35 unique malicious users present in the test partition, giving a user-level recall of 0.400000. This user-level result is substantially higher than the user-week-level recall of 0.114650, indicating that although the detector did not flag every malicious user-week belonging to a given user, it was considerably more likely to flag at least one user-week associated with a malicious user during the evaluation period.

3.9 Top-1% operating point

To examine detection behaviour independently of the frozen threshold, a Top-1% operating point was evaluated by selecting the highest-ranked 1 percent of ensemble scores on the test partition. This configuration produced 274 alerts, comprising 29 true positives and 245 false positives, for a precision of 0.105839, a recall of 0.184713, and an F1-score of 0.134571. At the user level, 20 of the 35 unique malicious test users were alerted at least once, giving a user-level recall of 0.571429.

Compared with the frozen threshold, the Top-1% operating point achieved substantially higher recall (0.184713 versus 0.114650) and user-level recall (0.571429 versus 0.400000), at the cost of lower precision (0.105839 versus 0.219512) and a considerably larger alert volume (274 versus 82 alerts). This comparison illustrates a practical precision-recall trade-off available to a security operations team. a volume-based operating rule, such as reviewing a fixed top percentage of weekly risk scores, can surface substantially more malicious user activity than the frozen scientific threshold, at the cost of a proportionally larger investigation workload (RQ5).

3.10 Two-week persistence analysis

A two-week persistence analysis examined whether requiring a user to be flagged in two consecutive weeks, rather than a single week, altered detection behaviour. This persistence rule produced 10 alerts, corresponding to 9 unique alerted users, with 8 true positives and 2 false positives, a precision of 0.800000, a recall of 0.050955, and an F1-score of 0.095808. At the user level, 8 of the 35 unique malicious test users were identified through persistent alerting, giving a user-level recall of 0.228571.

Among the three principal operating configurations examined (the frozen threshold, the Top-1% rule, and the persistence rule), the persistence rule produced the highest precision (0.800000), at the cost of the lowest recall (0.050955) and user-level recall (0.228571). This result demonstrates that requiring repeated anomalous behaviour across consecutive weeks substantially reduces false-positive volume, which may be valuable where analyst investigation capacity is limited, but that this benefit comes at the cost of missing malicious users whose anomalous behaviour was not sustained across two consecutive observation windows (RQ5).

3.11 Operational risk-tier alerting

Independently of the frozen scientific threshold, the ensemble risk scores were mapped to validation-derived percentile-based risk tiers (Critical, High, Medium, Low, Minimal), intended to support an operational security-operations-centre (SOC) alerting rule in which the High and Critical tiers trigger analyst attention. Table 6 reports the resulting test-partition tier distribution.

Table 6: Test-partition risk-tier distribution (27,302 user-weeks)

Risk tier	User-weeks	Share (%)
Critical	296	1.084
High	1,235	4.524
Medium	3,205	11.740
Low	4,676	17.128
Minimal	17,890	65.524

The combined High and Critical tiers accounted for 1,531 user-weeks on the test partition, an operational alert rate of 5.607648 percent. Across the complete prediction population (validation and test combined, 32,920 user-weeks), the expected and stored High-plus-Critical counts both equalled 1,812, with zero mismatches between the two calculations, confirming internal consistency of the risk-tier assignment pipeline.

It is important to distinguish the operational risk-tier alert rate (5.607648 percent, 1,531 test alerts) from the frozen scientific threshold alert rate reported in Section 3.8 (0.300344 percent, 82 test alerts). The two thresholds serve different purposes: the scientific threshold was selected to support the headline precision/recall/F1 evaluation reported against the 0.993402 cut-point, whereas the operational risk tiers were derived directly from the validation-period percentile distribution of ensemble scores and are intended to illustrate a plausible, considerably higher-volume SOC alerting configuration. These two thresholds must not be conflated when interpreting the results reported in this study.

3.12 Precision at fixed alert volumes

Table 7 reports Precision@K for the ensemble at four fixed alert volumes on the test partition.

Table 7: Precision@K for the weighted percentile-rank ensemble (test partition)

K	Precision@K
10	0.50
25	0.40
50	0.36
100	0.19

Precision@K declined monotonically as the alert volume K increased, from 0.50 at K=10 to 0.19 at K=100. This pattern is consistent with the ranking behaviour expected under a highly imbalanced evaluation population, the highest-ranked observations contain a disproportionately high concentration of malicious user-weeks, and precision decreases as progressively lower-ranked observations are included. The Precision@10 value of 0.50 indicates that, among the ten highest-ranked ensemble scores on the test partition, half corresponded to genuinely malicious user-weeks, providing evidence that the fused ranking was strongly informative at the extreme upper tail of the score distribution.

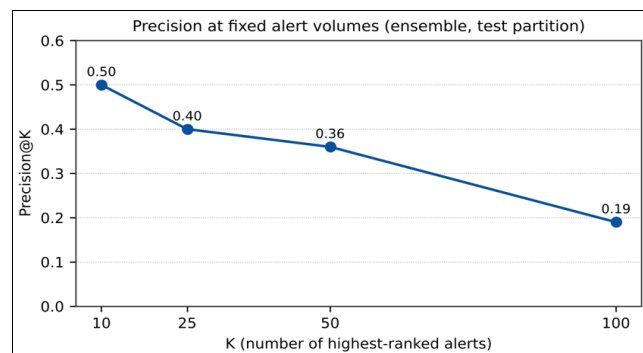


Fig 6: Precision at fixed alert volumes (ensemble, test partition)

3.13 Reproducibility verification

The deterministic behaviour of the fusion pipeline was independently verified through three separate experiment runs (A, B, and D), all of which produced an identical test PR-AUC of 0.09421108401734959. This value matched the authoritative PR-AUC computed directly from the stored Parquet artifacts (0.09421108401734961) and the corresponding ROC-AUC (0.8993543754758885) to the precision reported. A separate reconstruction check, in which the ensemble score was recomputed from the stored per-detector percentile scores and the frozen weights, produced a maximum absolute difference of zero, a mean absolute difference of zero, and zero nonzero differences relative to the stored ensemble score. These checks confirm that the reported results are exactly reproducible under the fixed random-state configuration (random_state=42) used throughout the study (RQ5).

3.14 Discussion

RQ1 (behavioural representation): The 86-feature user-week representation, incorporating base behavioural, temporal/multi-timescale, and week-on-week features across five activity modalities, provided sufficient discriminative information for all four unsupervised detectors to achieve ROC-AUC values well above chance level, and for the fused ensemble to achieve a test PR-AUC more than three times the base rate implied by the 157-in-27,302 class imbalance (0.00575). This indicates that the representation captured behavioural signal relevant to the documented malicious scenarios, although the comparatively low absolute PR-AUC values achieved by all configurations also confirm that ranking malicious user-weeks within a highly imbalanced population remains a difficult task.

RQ2 (comparative detector performance): The four detectors were not equally effective. COPOD, a statistical, distribution-based detector, achieved the strongest individual performance on both PR-AUC and ROC-AUC, while MiniBatch K-Means, a clustering-based detector, achieved the weakest performance on both measures. Isolation Forest and SGD One-Class SVM occupied an intermediate position. This ordering should be interpreted as evidence specific to the CERT r4.2 behavioural representation used in this study rather than as a general statement about the relative merit of isolation-based, clustering-based, boundary-based, and statistical anomaly detection, and is broadly consistent with prior reports that unsupervised ensembles and individual anomaly detectors can achieve useful but imperfect ranking performance on

CERT-derived insider-threat data [6, 12, 14]. The Spearman correlation analysis (Section 3.3) further showed that this performance ordering did not align with the detectors' pairwise ranking correlations: the weakest individual detector, MiniBatch K-Means, was among the most strongly correlated with the other detectors, while SGD One-Class SVM, of intermediate individual performance, was the least correlated with the remaining three, indicating that individual detector strength and ranking independence are distinct properties that must each be assessed separately when evaluating candidate detectors for a fused ensemble.

RQ3 (percentile-rank fusion compared with alternative normalisation strategies): Under identical frozen detector weights, replacing the percentile-rank transformation with raw-score or Min-Max normalisation reduced PR-AUC on both the validation and test partitions (Section 3.5), providing direct evidence that the percentile-rank transformation itself, rather than the detector-weight configuration alone, contributed to the ensemble's precision-recall ranking performance. This supports percentile-rank transformation as the preferable basis for fusing the four detectors' outputs in this setting when PR-AUC is the measure of interest. However, raw-score fusion achieved the highest test ROC-AUC of the three strategies (0.928916, against 0.899354 for percentile-rank fusion), so the advantage of percentile-rank transformation was specific to the higher-precision region of the ranking that PR-AUC emphasises rather than uniform across every ranking measure.

RQ4 (detector contribution): The ablation and weight-sensitivity analyses provide converging evidence that COPOD was the principal contributor to ensemble performance, that SGD One-Class SVM and Isolation Forest contributed comparatively modest and largely neutral effects, and that MiniBatch K-Means, under the frozen weight configuration, reduced rather than improved both PR-AUC and ROC-AUC relative to its exclusion. This finding illustrates that ensemble inclusion of a structurally diverse detector does not guarantee a complementary contribution [10, 11], and that empirical ablation is necessary to establish whether a given detector improves or degrades a fused configuration, rather than assuming diversity alone to be beneficial. The Spearman correlation analysis (Section 3.3) provides a mechanistic account of this result. MiniBatch K-Means was the detector most strongly correlated with Isolation Forest and COPOD ($\rho = 0.945$ and 0.832 respectively), indicating that it contributed a largely redundant rather than complementary ranking signal, whereas SGD One-Class SVM, the detector least correlated with the others ($\rho = 0.579$ - 0.629), provided a more independent contribution consistent with its modest but nonzero ablation effect on PR-AUC.

RQ5 (effect of operating configuration): The threshold-based, Top-1%, and two-week persistence operating configurations illustrate a precision-recall trade-off directly relevant to operational deployment. The frozen scientific threshold produced low alert volume and low recall, the Top-1% configuration substantially increased recall and user-level recall at the cost of alert volume and precision, and the two-week persistence rule achieved the highest precision at the lowest recall. The separate operational risk-tier alerting rule (High-plus-Critical tiers) produced a considerably higher test alert rate (5.607648 percent) than the frozen scientific threshold (0.300344 percent),

illustrating that the choice of operating rule has a substantial effect on the resulting analyst workload and should be selected according to the available investigation capacity of a given deployment context rather than treated as a fixed property of the detection model.

Limitations and threats to validity: Several limitations affect the interpretation of these results. First, the study relies on the CERT r4.2 simulated dataset, although it provides documented ground truth, the simulated nature of the underlying activity may not fully represent the behavioural diversity of a genuine organisational environment. Second, the frozen ensemble weights were selected using a single validation partition from a specific chronological window and may not generalise to datasets with different behavioural characteristics or class-imbalance ratios without re-validation. Third, the ablation analysis renormalised the existing frozen weights when a detector was removed, rather than independently reselecting weights for each reduced detector subset, the reported ablation results should therefore be interpreted as evidence of each detector's marginal contribution to the specific frozen configuration rather than as an exhaustive search for an optimal detector subset. Fourth, the engineering demonstration described briefly in Section 2 as providing operational context for the risk-tier alerting rule, including the FastAPI backend, React dashboard, and live simulation components, was not used to generate any of the scientific evaluation results reported in this section and should not be interpreted as independent empirical evidence. Finally, the comparatively low absolute precision at the frozen threshold indicates that, as with much insider-threat detection research under severe class imbalance, the resulting system is best understood as a prioritisation aid for analyst investigation rather than a mechanism for automated, high-confidence identification of malicious users.

4. Conclusion

This study investigated whether heterogeneous unsupervised anomaly detection algorithms, applied to a common multi-modal behavioural representation, could be combined through weighted percentile-rank score fusion to support insider-threat detection under severe class imbalance. Four structurally different detectors, Isolation Forest, MiniBatch K-Means, SGD One-Class SVM, and COPOD, were independently trained on an 86-feature user-week behavioural representation derived from the CERT Insider Threat Dataset Release 4.2, and their anomaly scores were transformed into empirical percentile ranks and combined using validation-selected, subsequently frozen weights.

On a held-out test partition of 27,302 user-weeks, containing 157 malicious user-weeks, the weighted ensemble achieved a PR-AUC of 0.094211 and a ROC-AUC of 0.899354, exceeding the PR-AUC of every individual detector while not uniformly exceeding every individual detector's ROC-AUC. A direct comparison against raw-score and validation-fitted Min-Max weighted fusion, using identical frozen detector weights, confirmed that the percentile-rank transformation itself contributed to this PR-AUC advantage, improving test PR-AUC by approximately 7.2 percent over raw-score fusion and 14.3 percent over Min-Max fusion, although raw-score fusion achieved the highest test ROC-AUC of the three strategies. COPOD was identified as the dominant contributor to ensemble performance, both as the strongest individual detector and through leave-one-

detector-out ablation, which showed that its removal reduced test PR-AUC by approximately 23.7 percent. Conversely, MiniBatch K-Means, the weakest individual detector, was found to reduce rather than improve ensemble performance under the frozen weight configuration, a result that would not have been apparent without the ablation and weight-sensitivity analyses conducted in this study. Threshold-based, Top-1 percent, and two-week persistence operating configurations further demonstrated a practical precision-recall trade-off, while a separate validation-derived risk-tier alerting rule illustrated a substantially higher-volume operational alternative to the frozen scientific threshold. All reported results were independently verified for exact reproducibility under a fixed random-state configuration.

The importance of this study lies in demonstrating that ensemble construction through percentile-rank fusion can improve PR-AUC relative to a heterogeneous set of individually trained unsupervised anomaly detectors, while simultaneously showing, through systematic ablation, that not every structurally diverse detector contributes positively to such a fusion. This distinction is relevant to insider-threat detection research and practice, where ensemble methods are frequently proposed without empirical confirmation that each constituent model contributes complementary rather than redundant or counter-productive information. The relevance of the study extends to the broader behavioural-analytics and anomaly-detection literature, in which percentile-rank and other rank-based fusion strategies offer a principled alternative to raw-score averaging when combining detectors whose outputs are not naturally comparable.

In summary, the study provides empirical evidence that weighted percentile-rank fusion of heterogeneous unsupervised anomaly detectors can provide a modest but measurable PR-AUC improvement over the best individual detector in a severely imbalanced insider-threat detection setting, and that the empirical contribution of each constituent detector should be verified through ablation and sensitivity analysis rather than assumed from algorithmic diversity alone. Further scope for this line of research includes extending the evaluation to additional behavioural datasets and organisational contexts, investigating adaptive or periodically re-validated ensemble weighting as behavioural populations evolve over time, and extending the ablation methodology to include full weight re-optimisation for each candidate detector subset rather than renormalisation of an existing frozen configuration.

5. Recommendations

Based on the findings of this study, the following recommendations are proposed for future research and for practitioners considering similar behavioural score-fusion approaches.

First, detector contribution should always be empirically verified through leave-one-detector-out ablation and weight-sensitivity analysis before a multi-detector ensemble is deployed or reported as beneficial. This study demonstrated that a structurally diverse detector, MiniBatch K-Means, reduced rather than improved ensemble performance under its frozen weight, a result that would not have been identified from aggregate performance figures alone.

Second, organisations adopting a similar framework should periodically re-validate ensemble weights using recent

behavioural data, rather than treating a single validation-derived weight configuration as permanently optimal, because behavioural populations and the statistical characteristics that make a detector such as COPOD effective may change over time.

Third, the operating configuration used for alerting, whether a fixed scientific threshold, a percentage-based ranking rule such as the Top-1 percent configuration, a persistence-based rule, or a percentile-derived risk tier, should be selected according to the available analyst investigation capacity of the specific deployment context, given the substantial differences in alert volume and precision-recall trade-off demonstrated across these configurations in this study.

Fourth, future work should extend the leave-one-detector-out methodology by independently re-optimising ensemble weights for each candidate detector subset, rather than renormalising the weights of an already-frozen configuration, in order to determine whether a reduced three-detector ensemble could achieve equal or superior performance to the evaluated four-detector configuration.

Fifth, future research should evaluate the weighted percentile-rank fusion methodology on additional insider-threat or behavioural-anomaly datasets beyond CERT r4.2, to assess whether the comparative ordering of detection paradigms and the dominant contribution of statistical, distribution-based detection observed in this study generalise to other organisational behavioural contexts.

Finally, given that the engineering demonstration components referenced in this study, including the analyst-facing dashboard and live simulation capability, were explicitly excluded from the scientific evaluation reported here, future work should pursue a formal, prospective operational evaluation of the framework's alerting behaviour under live or near-live behavioural telemetry, separate from the retrospective offline evaluation conducted in this study.

6. References

1. Cappelli D, Moore A, Trzeciak R. The CERT Guide to Insider Threats. How to Prevent, Detect, and Respond to Information Technology Crimes. Boston, Addison-Wesley, 2012, 1-400.
2. Homoliak I, Toffalini F, Guarnizo J, Elovici Y, Ochoa M. Insight into insiders and IT. A survey of insider threat taxonomies, analysis, defenses, and analysis methods. *ACM Computing Surveys*. 2019; 52(2):1-40.
3. Glasser J, Lindauer B. Bridging the gap. A pragmatic approach to generating insider threat data. *Proc IEEE Security and Privacy Workshops*, San Francisco, 2013, 98-104.
4. Gartner Inc. Market guide for user and entity behavior analytics. Gartner Research, 2017.
5. Chandola V, Banerjee A, Kumar V. Anomaly detection. A survey. *ACM Computing Surveys*. 2009; 41(3):1-58.
6. Le DC, Zincir-Heywood N. Anomaly detection for insider threats using unsupervised ensembles. *IEEE Transactions on Network and Service Management*. 2021; 18(2):1152-1164.
7. Li Z, Zhao Y, Botta N, Ionescu C, Hu X. COPOD. Copula-based outlier detection. *Proc IEEE International Conference on Data Mining*, Sorrento, 2020, 1118-1123.
8. Liu FT, Ting KM, Zhou ZH. Isolation forest. *Proc IEEE International Conference on Data Mining*, Pisa, 2008, 413-422.

9. Zhao Y, Nasrullah Z, Li Z. PyOD. A Python toolbox for scalable outlier detection. *Journal of Machine Learning Research*. 2019; 20(96):1-7.
10. Dietterich TG. Ensemble methods in machine learning. In. *Multiple Classifier Systems, Lecture Notes in Computer Science*, vol 1857. Berlin, Springer, 2000, 1-15.
11. Aggarwal CC, Sathe S. *Outlier Ensembles. An Introduction*. Cham, Springer, 2017, 1-277.
12. Bin Sarhan B, Altwaijry N. Insider threat detection using machine learning approach. *Applied Sciences*. 2023; 13(1):259.
13. Artioli P, Maci A, Magri A. A comprehensive investigation of clustering algorithms for user and entity behavior analytics. *Frontiers in Big Data*. 2024; 7:1375818.
14. Alzaabi FR, Mehmood A. A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods. *IEEE Access*. 2024; 12:30907-30927.
15. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. *Proc 23rd International Conference on Machine Learning*, Pittsburgh, 2006, 233-240.
16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* Scikit-learn. Machine learning in Python. *Journal of Machine Learning Research*. 2011; 12:2825-2830.
17. Sculley D. Web-scale k-means clustering. *Proceedings of the 19th International Conference on World Wide Web*, 2010, 1177-1183.
18. Scholkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. *Neural Computation*. 2001; 13(7):1443-1471.