Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 10, pp 50-58, October-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.310006
Available Online at https://www.jaiet.com

A Four-Algorithm Unsupervised Voting Ensemble for Debit Card Fraud Detection: Alert Burden Reduction and Threshold Calibration

[1]*Kufre Ambrose Edet, [2]Patience U. Usip

[1]Dept. of Information Systems and Technology, Faculty of Computing, National Open University of Nigeria (NOUN), Abuja, Nigeria
[2]Department of Computer Science, University of Uyo, Uyo, Akwa Ibom State, Nigeria
*Corresponding Author's E-mail: nou255316957@noun.edu.ng

Abstract: Fraud detection in Nigerian deposit money banks is constrained by the near-total absence of labelled fraud repositories, which excludes supervised learning for most institutions and leaves unsupervised anomaly detection as the only deployable alternative. Yet no study has evaluated a heterogeneous four-algorithm voting ensemble on a card-fraud benchmark, or quantified the alert-burden consequences of consensus voting. Isolation Forest, Local Outlier Factor, One-Class Support Vector Machine and a deep Autoencoder were trained exclusively on legitimate transactions and their binary votes combined by a three-of-four majority rule. On a held-out partition of 42,559 transactions containing 71 fraud cases, the ensemble attained the highest F1-score (0.2437), the highest precision (0.1566) and the lowest false positive rate (0.0049) of all five models evaluated, reducing false alerts by 90.3% relative to the highest-recall constituent and lowering the false alarm ratio from 36.8:1 to 5.4:1. McNemar testing confirmed the F1 advantage as significant against three of the four constituents, with bootstrap confidence intervals excluding zero for all four comparisons. Both the precision advantage and the alert-burden reduction replicated on 831,123 African mobile-money transactions, where the F1 advantage was significant against all four constituents at p below 0.001. Threshold ablation across all four consensus rules showed that three-of-four is F1-optimal on the card benchmark but not on the mobile-money corpus, where unanimity achieved an F1-score of 0.3749. The optimal consensus threshold is therefore a property of the deployment data rather than of the architecture, which argues for per-deployment calibration rather than transfer of a published value.

Keywords: Anomaly detection, Ensemble learning, Credit card fraud, Class imbalance, Threshold calibration.

I. INTRODUCTION

Nigerian debit card fraud inflicted losses exceeding N59.5 billion in 2023, while electronic transactions processed in the first half of 2024 reached NGN 1.56 quadrillion and reported fraud cases rose by 468% over the same period [1], [2]. Detection in Nigerian deposit money banks nonetheless remains predominantly rule-based [3], [4]. Rule sets are transparent and auditable but brittle: they cannot generalise to fraud typologies absent from their specification, and they degrade as fraud actors adapt [5], [6].

Machine learning offers an alternative, but the dominant paradigm in the literature is supervised. Supervised models achieve high F1-scores on benchmark corpora, yet they presuppose a labelled historical fraud repository, which Alaba *et al.* [3] document that most Nigerian deposit money banks do not maintain. Ayodeji [4], interviewing fifteen information technology leaders across Nigerian financial institutions, found detection strategies predominantly reactive and analytically shallow. Unsupervised anomaly detection requires only legitimate transactions for training and is therefore deployable at institutions lacking fraud labels. This is not a marginal advantage: it is the difference between a system deployable at all twenty-four licensed deposit money banks and one deployable at perhaps five.

II. PROBLEM DEFINITION

Two gaps follow. First, no published study evaluates a heterogeneous four-algorithm unsupervised voting ensemble for card fraud detection. Prior ensembles pair two algorithms or combine supervised with unsupervised components: Adejoh *et al.* [7] combine autoencoders with self-organising maps, Asuai *et al.* [8] pair an artificial neural network with gradient boosting, Ajagbe *et al.* [9] compare detectors without combining them, and Jejeniwa *et al.* [10] hybridise a graph neural network with Isolation Forest. Second, the literature reports ensemble gains in

© 2026 JAIET All Rights Reserved CC BY 4.0 License

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 10, pp 50-58, October-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.310006
Available Online at https://www.jaiet.com

F1 without examining the alert-burden consequence of the aggregation mechanism. For an investigation unit with finite capacity, the ratio of false alerts to genuine detections determines whether a system's output can be processed at all, and that ratio is a direct function of the voting rule.

This paper makes three contributions. It presents, to our knowledge, the first four-algorithm unsupervised voting ensemble evaluated on the ULB card-fraud benchmark, and quantifies its advantage over each constituent on F1, precision and false positive rate with McNemar and bootstrap confirmation. It measures the alert-burden reduction that consensus voting purchases, and the recall it costs, against every constituent. It tests whether these findings are properties of the architecture or artefacts of one corpus by replication on a structurally dissimilar African mobile-money dataset, and reports that the F1-optimal consensus threshold differs between the two — which establishes that threshold selection is a deployment decision, not an architectural constant.

III. LITERATURE SURVEY

3.1 Unsupervised anomaly detection

Anomaly detection identifies observations deviating from an established notion of normality without requiring labelled examples of the anomalous class. Four paradigms are relevant here, and each encodes a different definition of abnormality.

Isolation Forest isolates anomalies by recursive random partitioning, on the principle that anomalies are isolated at shallower tree depths than normal points [11]; its complexity is approximately linear in sample size, so it scales to high-volume transaction streams. Local Outlier Factor compares a point's local density with that of its nearest neighbours, detecting anomalies that are locally rather than globally deviant [12], which suits account-specific behavioural departures. One-Class Support Vector Machine learns a kernel-space hyperplane enclosing the bulk of normal training instances [13]; it is sensitive to its ν and γ hyperparameters and its training complexity is quadratic in sample size. Autoencoders learn a compressed representation of normal data and reconstruct it, flagging observations whose reconstruction error exceeds a threshold [14].

The four target global isolation, local density, support boundary and reconstruction fidelity respectively. Their errors are plausibly weakly correlated, which is the precondition for ensemble gain.

3.2 Ensembles in fraud detection

Ensemble superiority is consistently reported. Zavvar *et al.* [15] combine synthetic minority oversampling, an autoencoder, a convolutional neural network and an attention mechanism to reach F1 = 0.9060 on the ULB dataset. Asuai *et al.* [8] report gains from an artificial neural network and gradient boosting hybrid. Jejeniwa *et al.* [10] report F1 = 0.92 and AUC-ROC = 0.98 from a graph neural network and Isolation Forest hybrid on real Nigerian financial technology data, the highest documented performance on Nigerian transaction data in the literature reviewed. Ajagbe *et al.* [9], evaluating algorithms in an African institutional setting, find XGBoost dominant in supervised settings while noting that Isolation Forest and support vector methods supply complementary anomaly signal. Adejoh *et al.* [7] construct an adaptive unsupervised ensemble of autoencoders and self-organising maps, reporting accuracy 0.980 and F1 0.967 on benchmark data.

Two observations follow. The strongest reported results are supervised and therefore not deployable without labels. And no study reports a four-algorithm unsupervised vote, nor measures the alert burden that the voting rule produces.

3.3 Evaluation under extreme class imbalance

Accuracy is uninformative when fraud prevalence is below 0.2%: a classifier assigning every case to the majority class attains 99.83% accuracy on the ULB corpus while detecting nothing [16]. Precision, recall, F1 and false positive rate are the operative metrics, and the false positive rate governs investigator workload because it determines alert volume [3]. Fawcett [17] sets out the construction and interpretation of receiver operating characteristic curves, and notes that the area beneath the curve summarises ranking quality across all thresholds — a property that presupposes a continuous decision score.

Class imbalance handling is contested. Chawla *et al.* [18] introduced synthetic minority oversampling; Yu *et al.* [19] compare sampling strategies and ensemble methods. Because oversampling applied before partitioning leaks synthetic minority structure into the evaluation set, we restrict it to the training partition. Btoush *et al.* [20], reviewing 181 articles, identify class imbalance and concept drift as the persistent unresolved challenges of the field; our contribution addresses the first in the unsupervised setting.

© 2026 JAIET All Rights Reserved

CC BY 4.0 License

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 10, pp 50-58, October-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.310006
Available Online at https://www.jaiet.com

3.4 Nigerian context

Onyeama [21] applies autoencoders to Nigerian credit card data and finds them superior to principal component analysis. Ayodeji [4] and Alaba *et al.* [3] supply the institutional context: the former documenting reactive, analytically shallow detection practice, the latter quantifying adoption determinants and identifying cost, competence and compliance as barriers. Alaba *et al.* [3] find that only the five largest banks show evidence of partial artificial intelligence deployment, while the remaining nineteen licensed deposit money banks rely predominantly on rule-based or manual approaches. A method that performs respectably without labels, at institutions that have none, addresses a real operational gap.

IV. METHODOLOGY

4.1 Datasets

The primary corpus is the ULB card-fraud benchmark: 284,807 anonymised European card transactions of which 492 (0.173%) are fraudulent, with 28 principal components plus transaction amount and elapsed time [22]. We removed 1,081 exact duplicate records before partitioning, reducing the corpus to 283,726 transactions, because a record present in both training and evaluation partitions allows a model to appear to generalise where it has memorised. Stratified 70/15/15 partitioning produced 198,608 training records, 42,559 validation records and a held-out test partition of 42,559 transactions containing 71 fraud cases (0.172% prevalence, an imbalance of 578:1). All reported metrics derive from that test partition, which was not used at any stage of fitting or threshold selection.

Cross-dataset validation used PaySim, an agent-based simulation of African mobile-money transactions calibrated to real transaction logs [23], selected because its fraud typologies resemble the account-draining patterns documented in the Nigerian context and because it differs from the card benchmark in scale, domain, feature composition and prevalence. Of 6,362,620 raw records, 8,213 (0.129%) are fraudulent, and fraud occurs exclusively within the CASH_OUT and TRANSFER types. Filtering to those types leaves 2,770,409 transactions; a stratified 70/30 partition produced a held-out test partition of 831,123 transactions containing 2,464 fraud cases (0.296% prevalence).

4.2 Preprocessing

For the card benchmark, transaction amount was log-transformed as log(1 + Amount) to reduce extreme right skew, because untransformed amounts distort the distance and kernel metrics on which Local Outlier Factor and One-Class SVM depend. Elapsed time was decomposed into hour of day. Thirty features were supplied to every detector: V1–V28, Amount_log and Hour_of_Day, standardised using statistics fitted on the training partition only. Synthetic minority oversampling was applied to the training partition only, raising training fraud instances from 331 to 19,827 [18].

For the mobile-money corpus, twelve features were engineered from the raw fields: Amount_log, Hour_of_Day, a binary transfer indicator, originating and destination balances before and after the transaction, the two balance changes, a balance-change ratio, and two binary indicators marking an originating balance of zero after the transaction and a destination balance of zero before it, which capture the account-draining signature of fraudulent cash-out. The four detectors were fitted on a random subsample of 60,000 legitimate training transactions, because One-Class SVM training is intractable on the full 1.93 million legitimate training records. This subsampling affects model fitting only and does not enter any reported metric.

4.3 Detectors

All four detectors were fitted on legitimate transactions only; no fraud label was used at any stage of fitting.

- **Isolation Forest:** 200 estimators, contamination 0.005, seed 42, no bootstrap.
- **Local Outlier Factor:** 20 neighbours, contamination 0.005, novelty mode enabled for test-time inference.
- **One-Class SVM:** radial basis function kernel, $\nu = 0.01$, γ set to scale.
- **Autoencoder:** $30 \rightarrow 64 \rightarrow 32 \rightarrow 64 \rightarrow 30$, rectified linear hidden activations, sigmoid output, mean squared error loss, Adam at learning rate 0.001, batch size 256, maximum 50 epochs, early stopping with patience 5. The decision threshold is set at the 95th percentile of training reconstruction error [7].

Implementation used Python 3.10, scikit-learn [24] and Keras. Seeds were fixed at 42; re-execution on the same data reproduces every reported figure.

4.4 Ensemble rule and evaluation protocol

Each detector emits a binary vote. A transaction is flagged when the vote sum reaches at least three of four. The threshold was chosen over simple majority to prioritise precision,

© 2026 JAIET All Rights Reserved CC BY 4.0 License

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 10, pp 50-58, October-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.310006
Available Online at https://www.jaiet.com

consistent with the operational priorities of Nigerian investigation teams [3], [16], on the reasoning that requiring three independent paradigms to agree reduces the probability that an idiosyncratic false positive from any single detector propagates to the output. Section 5.3 reports the ablation across all four possible thresholds on both corpora.

Six metrics were computed for all five models: accuracy, precision, recall, F1, false positive rate and area under the receiver operating characteristic curve. Accuracy is reported for completeness and is not used as an evaluation criterion, for the reason given in Section 3.3. We report false alarm ratio — false alerts raised per genuine detection — alongside the standard metrics, because it is the quantity that determines investigative workload.

Comparative claims are supported by McNemar's test with Yates continuity correction on discordant pairs, using the exact binomial variant where the discordant count falls below twenty-five, and by percentile bootstrap confidence intervals for the difference in F1 with 2,000 resamples. The significance level was 0.05 throughout.

V. RESULTS

5.1 Individual detectors

Table 1 reports performance on the held-out card partition. The detectors fail in different directions, which is the empirical basis for combining them. One-Class SVM and the Autoencoder achieve high recall (0.8169 and 0.8310) but low precision (0.0908 and 0.0265); the Autoencoder raised 2,171 false alerts against 59 genuine detections. Isolation Forest is balanced but moderate. Local Outlier Factor is the weakest on this corpus (F1 = 0.0610), consistent with the known degradation of density-based methods under extreme imbalance, since at 578:1 the local neighbourhood is dominated by legitimate transactions [9].

Table 1: Performance on the ULB held-out partition
(n = 42,559; 71 fraud; 42,488 legitimate)

Model	F1	Precision	Recall	FPR	TP	FP	False alarm ratio
Isolation Forest	0.1889	0.1176	0.4789	0.0060	34	255	7.5 : 1
Local Outlier Factor	0.0610	0.0402	0.1268	0.0051	9	215	23.9 : 1
One-Class SVM	0.1634	0.0908	0.8169	0.0137	58	581	10.0 : 1
Autoencoder	0.0513	0.0265	0.8310	0.0511	59	2,171	36.8 : 1
Ensemble (3-of-4)	**0.2437**	**0.1566**	0.5493	**0.0049**	39	210	**5.4 : 1**

AUC-ROC is not reported for comparison. Because the ensemble's decision score is an integer vote count admitting five values, its ROC curve has at most five operating points, and the area beneath it is not commensurable with that of a continuous anomaly score [17]. A degenerate majority-class classifier attains accuracy 0.99833 on this partition while detecting no fraud.

5.2 Ensemble performance and alert burden

The ensemble achieves the highest F1 (0.2437), the highest precision (0.1566) and the lowest false positive rate (0.0049) of all five models. Relative to the highest-recall constituent it reduces false alerts from 2,171 to 210, a reduction of 90.3%, and lowers the false alarm ratio from 36.8:1 to 5.4:1. For an investigation unit with finite capacity, this is the operative result: at 36.8:1, an investigator processes roughly thirty-seven false alerts for each genuine case surfaced; at 5.4:1, roughly five.

The cost is recall. The ensemble detects 39 of 71 fraud cases (0.5493), against 59 for the Autoencoder and 58 for One-Class SVM. This is the expected consequence of the consensus threshold: cases detected by fewer than three paradigms are not flagged, which means the 32 missed cases were ones about which at least two independent definitions of abnormality disagreed.

5.3 Threshold ablation

Of 249 flagged transactions, 240 received three votes and 9 received four. The near-absence of unanimity on this partition is the reason the four-of-four rule is unusable here.

© 2026 JAIET All Rights Reserved

CC BY 4.0 License

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 10, pp 50-58, October-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.310006
Available Online at https://www.jaiet.com

Table 2: Threshold ablation across both corpora

Voting rule	ULB TP	ULB FP	ULB F1	ULB FPR	PaySim TP	PaySim FP	PaySim F1	PaySim precision
≥1 of 4	59	2,376	0.0471	0.0559	1,679	45,069	0.0682	0.0359
≥2 of 4	58	631	0.1526	0.0149	1,448	9,479	0.2163	0.1325
≥3 of 4	39	210	**0.2437**	0.0049	987	3,695	0.2762	0.2108
≥4 of 4	4	5	0.1000	0.0001	591	98	**0.3749**	0.8578

On the card benchmark, three-of-four maximises F1. One-of-four maximises recall at an operationally unacceptable false positive rate of 0.0559; two-of-four still raises 631 false alerts; four-of-four achieves the highest precision (0.4444) but detects only 4 of 71 cases. On the mobile-money corpus the ranking changes: four-of-four achieves F1 = 0.3749 at precision 0.8578 with only 98 false alerts. The explanation is the frequency of unanimity — nine transactions on the card benchmark attracted four votes, against 689 on the mobile-money corpus. Greater separability in the simulated data allows all four detectors to agree far more often, and unanimity therefore ceases to suppress recall.

The optimal consensus threshold is a function of detector agreement in the deployment data, not a fixed architectural constant.

5.4 Statistical significance

McNemar's test confirms the F1 advantage as significant against three constituents (Table 3). Against Local Outlier Factor the result is not significant ($\chi^2 = 2.62$, $p = 0.1054$), though the bootstrap interval excludes zero ($\Delta F1 = +0.1826$, 95% CI [0.1253, 0.2394]). We report this as a non-significant result rather than claiming a fourth advantage: the correct interpretation of a non-significant paired test with a large point estimate is that the evidence is insufficient, not that the effect is absent.

Table 3: McNemar's test and bootstrap confidence intervals, ensemble versus each constituent (ULB)

Ensemble versus	χ^2	p	ΔF1	95% CI
Isolation Forest	44.46	< 0.001	+0.0547	[+0.0313, +0.0824]
Local Outlier Factor	2.62	0.1054	+0.1826	[+0.1253, +0.2394]
One-Class SVM	315.90	< 0.001	+0.0800	[+0.0350, +0.1216]
Autoencoder	1,899.85	< 0.001	+0.1917	[+0.1354, +0.2463]

5.5 Cross-dataset validation

Table 4 reports performance on the mobile-money corpus. All three advantages transfer: the ensemble again achieves the highest F1 (0.2762), the highest precision (0.2108) and the lowest false positive rate (0.0045), reducing false alerts by 90.8% relative to the Autoencoder, closely mirroring the 90.3% reduction on the card benchmark. McNemar testing confirms the F1 advantage against all four constituents at $p < 0.001$, with bootstrap intervals excluding zero in all four comparisons. The near-identical alert-burden reduction across corpora differing in scale, domain and prevalence indicates that this benefit is a structural property of consensus voting rather than an artefact of one dataset.

Table 4: Performance on the PaySim held-out partition
(n = 831,123; 2,464 fraud)

MODEL	F1	Precision	Recall	FPR	TP	FP	False alarm ratio
Isolation Forest	0.2311	0.1699	0.3608	0.0052	889	4,342	4.9 : 1

© 2026 JAIET All Rights Reserved

CC BY 4.0 License

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 10, pp 50-58, October-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.310006
Available Online at https://www.jaiet.com

MODEL	F1	Precision	Recall	FPR	TP	FP	False alarm ratio
Local Outlier Factor	0.1931	0.1442	0.2922	0.0052	720	4,273	5.9 : 1
One-Class SVM	0.2349	0.1434	0.6477	0.0115	1,596	9,531	6.0 : 1
Autoencoder	0.0679	0.0360	0.6088	0.0485	1,500	40,195	26.8 : 1
Ensemble (3-of-4)	**0.2762**	**0.2108**	0.4006	**0.0045**	987	3,695	**3.7 : 1**

VI. DISCUSSION

6.1 Alert burden as the operative constraint

The central finding is that consensus voting purchases a large reduction in investigator workload at a quantified cost in recall. The ensemble raised 210 false alerts against 39 genuine detections on the card benchmark and 3,695 against 987 on the mobile-money corpus, reductions of 90.3% and 90.8% relative to the highest-recall constituent. A model raising 2,171 false alerts for 59 genuine detections will produce the alert fatigue that Enehikhare and Odumuyiwa [16] identify as a primary driver of investigative quality degradation, and that Alaba *et al.* [3] observe institutionally as a source of scepticism toward artificial intelligence adoption. The ensemble's 5.4:1 ratio is a sevenfold reduction in that burden.

The lost recall should not be minimised. Thirty-two fraud cases were not flagged on the card benchmark. Those cases were detected by fewer than three paradigms, which means the evidence against them was weak by each of the four definitions of abnormality the system employs. A bank deploying this system would route them through rule-based controls, and the system should be positioned as an anomaly scoring layer augmenting, not replacing, existing rules — consistent with the framing of detection as an arms race rather than a static optimisation problem [6].

6.2 Threshold calibration is deployment-specific

The most practically consequential result is that the F1-optimal consensus threshold differs between the two corpora. On the card benchmark, three-of-four maximises F1; on the mobile-money corpus, unanimity does. The mechanism is detector agreement: nine transactions attracted four votes on one corpus, 689 on the other. A threshold tuned on one dataset and transferred to another would be mis-specified, and the mis-specification is not detectable from the published metric alone.

The design implication is that institutions deploying a consensus ensemble should calibrate the voting threshold on their own validation data rather than adopting a published value, and should report the calibration alongside the resulting performance. This is a low-cost procedure — the ablation in Table 2 requires nothing beyond the fitted detectors and a held-out partition — and it removes a source of silent error in cross-institutional transfer.

6.3 On the comparison of vote-based ensembles by AUC

We do not report an AUC comparison, and the reason is methodological rather than evidential. Area under the ROC curve summarises ranking quality across all thresholds, and that summary presupposes a decision score with many possible values [17]. A hard-voting ensemble's decision score is an integer count admitting k + 1 values for k detectors, so its ROC curve has at most five operating points here, and the area beneath it cannot be compared with that of a continuous anomaly score without comparing a step function to a continuous curve. Reporting an AUC deficit for such an ensemble would describe the arithmetic of the aggregation rule rather than a difference in discriminative ability.

The practical consequence is that vote-based ensembles should be evaluated on the metrics their decision rule produces — F1, precision, recall, false positive rate and false alarm ratio at the chosen threshold — and that comparisons with continuous-score models should be made at matched operating points. This is not a limitation of the present study but a property of hard voting that applies to any ensemble of this form.

6.4 Relation to prior benchmarks

Direct comparison with supervised results is methodologically inappropriate. Jejeniwa *et al.* [10] report F1 = 0.92 and AUC-ROC = 0.98 from a supervised graph neural network and Isolation Forest hybrid on real Nigerian financial technology data; Zavvar *et al.* [15] report F1 = 0.9060 from a supervised deep hybrid on the card benchmark. Both presuppose labelled fraud at training time.

Our F1 of 0.2437 is the highest unsupervised result among the models evaluated here, and the appropriate baseline is the individual unsupervised range of 0.0513 to 0.1889, against which

© 2026 JAIET All Rights Reserved CC BY 4.0 License

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 10, pp 50-58, October-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.310006
Available Online at https://www.jaiet.com

the ensemble represents an improvement of 29% to 375%. Adejoh *et al.* [7], the most closely comparable unsupervised ensemble study, report F1 = 0.967 and accuracy 0.980; the difference is large and attributable in part to their adaptive thresholding and in part to differences in partitioning and evaluation protocol that cannot be reconciled from the published account. We make no claim to superiority over that study. The gap between supervised performance and our unsupervised result is not a failure of architecture; it is the measured cost of operating without labels, which is the condition of nineteen of Nigeria's twenty-four licensed deposit money banks [3].

6.5 Limitations

First, both corpora are proxies. The card benchmark is European with principal-component anonymised features, so no substantive interpretation of individual predictors is possible, and PaySim is simulated. Neither carries Nigerian transaction signatures such as bank verification number linkage, USSD channel indicators or point-of-sale terminal metadata [21].

Second, the Autoencoder uses a sigmoid output activation while its inputs are standardised and therefore unbounded, whereas a sigmoid is bounded to [0, 1]. The output layer cannot fully reconstruct its input and reconstruction error is correspondingly inflated. The detector remains functional because its threshold is derived from the same error distribution, but its reconstruction quality is constrained; a linear output activation is the appropriate specification and would likely raise both its individual performance and the ensemble's recall ceiling.

Third, Local Outlier Factor's weak individual performance (F1 = 0.0610) suggests that the ensemble's upper bound is set by its weakest member. Replacement, not reweighting, is the appropriate remedy where a stronger detector is available.

Fourth, the study evaluates detection performance only. No user acceptance study was conducted with practising investigators and no live banking deployment was attempted, so the operational interpretation in Section 6.1 is an inference from the metrics and from the published institutional literature rather than a measured outcome.

VII. CONCLUSION

A four-algorithm unsupervised voting ensemble of Isolation Forest, Local Outlier Factor, One-Class Support Vector Machine and Autoencoder, combined by a three-of-four majority rule, achieves the highest F1-score, the highest precision and the lowest false positive rate of all five models evaluated on the ULB card-fraud benchmark. It reduces false alerts by 90.3% relative to the highest-recall constituent, lowering the false alarm ratio from 36.8:1 to 5.4:1, at the cost of detecting 39 of 71 fraud cases rather than 59. These advantages are statistically significant against three of four constituents and replicate on an independent African mobile-money corpus, where they are significant against all four and the alert-burden reduction is almost identical at 90.8%.

The optimal consensus threshold proved to be dataset-dependent: three-of-four maximises F1 on the card benchmark, unanimity on the mobile-money corpus. That result establishes that threshold selection is a deployment decision requiring local calibration, not an architectural constant to be transferred.

VIII. FUTURE SCOPE

Two directions follow: correcting the Autoencoder's output activation and replacing Local Outlier Factor with a stronger detector would raise the recall ceiling, and the architecture should be evaluated on genuine Nigerian transaction data in partnership with a settlement infrastructure provider or a willing deposit money bank, since the benchmark result reported here is best read as a lower bound.

DECLARATIONS

Conflict of interest: The authors declare no conflict of interest.

Funding: This research received no external funding.

Ethics: Not applicable; only publicly available, anonymised or simulated datasets were used, and no human participants or personally identifiable financial records were involved.

Data availability: The ULB credit card fraud dataset is publicly available from the Kaggle repository maintained by the Machine Learning Group at the Université Libre de Bruxelles (mlg-ulb/creditcardfraud). The PaySim mobile-money dataset is publicly available from the Kaggle repository maintained by Lopez-Rojas and colleagues (ealaxi/paysim1). The experimental script and the transaction-level prediction vectors from which all statistical tests were computed are available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

Kufre Ambrose Edet: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization.

© 2026 JAIET All Rights Reserved CC BY 4.0 License

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 10, pp 50-58, October-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.310006
Available Online at https://www.jaiet.com

Patience U. Usip: Supervision, Writing – review & editing, Project administration.

ACKNOWLEDGEMENT

The authors thank Kubiat Akpan Ekpo, Department of Information Systems and Technology, Faculty of Computing, National Open University of Nigeria, for comments on an earlier draft and for technical assistance during the experimental runs. The first author thanks the Department for institutional support.

REFERENCES

[1] A. Green, "AI-driven financial intelligence systems: A new era of risk detection and strategic analysis," *Applied Sciences in Engineering and Technology*, vol. 3, no. 1, pp. 1–22, 2024.

[2] Nigeria Inter-Bank Settlement System, Nigeria Fraud Report H1 2024: Electronic Fraud Statistics for Nigerian Banking Sector. *NIBSS Fraud Desk,* 2024. [Online]. Available: https://www.nibss-plc.com.ng

[3] J. S. Alaba, S. J. Ahmed, A. P. Farida, and O. V. Oluwatosin, "Adoption of AI-driven fraud detection system in the Nigerian banking sector: An analysis of cost, compliance, and competency," *Economic Review of Nepal*, vol. 8, no. 1, pp. 16–33, 2025, doi: 10.3126/ern.v8i1.80740.

[4] I. A. Ayodeji, "Fraud detection and prevention in the Nigerian financial industry," *Ph.D. dissertation, Walden University*, 2024. [Online]. Available: https://scholarworks.waldenu.edu/dissertations/

[5] O. S. Adebayo, T. A. Favour-Bethy, O. Otasowie, and O. A. Okunola, "Comparative review of credit card fraud detection using machine learning and concept drift techniques," *International Journal of Computer Science*, vol. 12, no. 2, pp. 24–48, 2023.

[6] C. D. Ikemefuna, O. Okusi, A. C. Iwu, and S. Yusuf, "Adaptive fraud detection systems: Using ML to identify and respond to evolving financial threats," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 6, no. 9, pp. 1727–1735, 2024.

[7] J. Adejoh, N. Owoh, M. Ashawa, S. Hosseinzadeh, A. Shahrabi, and S. Mohamed, "An adaptive unsupervised learning approach for credit card fraud detection," *Machines*, vol. 13, no. 8, p. 651, 2025, doi: 10.3390/machines13080651.

[8] C. Asuai, O. K. Nana, and I. E. Destiny, Credit Card Fraud Detection Optimised Approach with Artificial Neural Networks and Gradient Boosting Model. *Delta State University, Abraka*, 2024.

[9] S. A. Ajagbe, S. Majola, and P. Mudali, "Comparative analysis of machine learning algorithms for money laundering detection," *Frontiers in Computer Science*, vol. 6, Art. no. 1376501, 2024, doi: 10.3389/fcomp.2024.1376501.

[10] T. O. Jejeniwa, T. O. Jejeniwa, and O. S. Owolabi, "Leveraging hybrid AI for real-time fraud detection: A case study on the efficacy of graph neural networks and anomaly detection in Nigerian FinTechs," *Journal of Digital Security and Forensics*, vol. 2, no. 2, pp. 137–143, 2025, doi: 10.29121/DigiSecForensics.v2.i2.2025.60.

[11] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," *in Proc. 8th IEEE Int. Conf. Data Mining, IEEE,* 2008, pp. 413–422.

[12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93–104, 2000.

[13] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[14] I.Goodfellow, Y. Bengio, and A. Courville, Deep Learning. *MIT Press*, 2016.

[15] M. Zavvar *et al.*, "Credit card fraud detection using deep learning: A hybrid SMOTE, autoencoder, convolutional neural network, and attention mechanism for credit card fraud detection," *Journal of Big Data*, vol. 13, Art. no. 21, 2026, doi: 10.1186/s40537-025-01331-2.

[16] C. C. Enehikhare and V. Odumuyiwa, "Comparative analysis of machine learning models for fraud detection in imbalanced credit card transaction datasets," *University of Ibadan Journal of Science and Logics in ICT Research*, vol. 14, no. 1, pp. 25–37, 2025.

[17] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[19] L. Yu, C. T. Chang, V. N. Huynh, S. Jesus, and J. Luo, "Credit card fraud detection with imbalanced datasets: A comparative study of sampling strategies and ensemble methods," *Applied Sciences*, vol. 13, no. 14, Art. no. 8372, 2023, doi: 10.3390/app13148372.

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 10, pp 50-58, October-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.310006
Available Online at https://www.jaiet.com

[20] E. A. L. M. Btoush, X. Zhou, R. Gururajan, K. C. Chan, R. Genrich, and P. Sankaran, "A systematic review of literature on credit card cyber fraud detection using machine and deep learning," *Peer J Computer Science*, vol. 9, Art. no. e1278, 2023, doi: 10.7717/peerj-cs.1278.

[21] F. C. Onyeama, "Unsupervised TensorFlow-based anomaly detection for Nigerian credit card fraud: Autoencoder vs. principal component analysis," *Journal of Computer Science and Its Applications*, vol. 29, no. 1, pp. 14–23, 2022.

[22] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," *Journal of Machine Learning Research*, vol. 22, no. 4, pp. 1–40, 2021.

[23] E. A. Lopez-Rojas, S. Ellegaard, and S. Axelsson, "PaySim: A financial mobile money simulator for fraud detection," *Simulation Modelling Practice and Theory*, vol. 108, Art. no. 102265, 2021, doi: 10.1016/j.simpat.2021.102265.

[24] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

AUTHORS BIOGRAPHY



Kufre Ambrose Edet is an Information Technology professional, educator, and researcher pursuing a Master of Science degree in Information Technology at the National Open University of Nigeria. His professional experience includes computer education, ICT training, and technology support. His research interests include artificial intelligence, explainable machine learning, financial fraud detection, and decision support systems, particularly their application to practical challenges in Nigeria.



Prof. Patience Usoro Usip is a Computer Science academic at the University of Uyo, Akwa Ibom State, Nigeria, where she serves as Head of the Department of Computer Science. Her research specialisation is Knowledge Representation and Reasoning, with particular interests in knowledge graphs, ontologies, Semantic Web technologies, and intelligent knowledge management systems. Her scholarly contributions include research on semantic knowledge retrieval, ontology-based systems, multilingual knowledge representation, and machine learning applications.

Citation of this Article:

Kufre Ambrose Edet, & Patience U. Usip. (2026). A Four-Algorithm Unsupervised Voting Ensemble for Debit Card Fraud Detection: Alert Burden Reduction and Threshold Calibration. *Journal of Artificial Intelligence and Emerging Technologies (JAIET)*. 3(10), 50-58. Article DOI: https://doi.org/10.47001/JAIET/2026.310006

***** End of the Article *****