Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 24 July 2026
Revised 18 Aug 2026
Accepted 06 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (22s): 330–347

KEYWORDS:
cybersecurity; inference attacks; privacy-preserving machine learning; behavioural analytics; differential privacy; k-anonymity.

Threat-Aware Security Evaluation of Privacy Controls Against Inference Attacks in Financial Behaviour Analytics

Nilesh Mali[1,2]*, Vidyullata Devmane[1]

[1]Department of Computer Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, India.
[2]Department of Information Technology, Vasantdada Patil Pratishthan's College of Engineering and Visual Arts, Mumbai, India.

*Corresponding author: nilesh.mali24@sakec.ac.in; nileshmali@pvppcoe.ac.in

ABSTRACT: Machine-learning analysis of financial behaviour offers valuable insight into financial well-being, but such models must also be protected against record linkage, attribute inference and model-output attacks. This study presents a threat-aware privacy–utility benchmark that compares eight privacy-control configurations on 52,034 respondents from the 2021 and 2024 National Financial Capability Study. A seven-item financial well-being (FWB) outcome was constructed independently of 47 behavioural predictors, ensuring a non-circular prediction task. Logistic regression, random forest, gradient boosting and support-vector machine classifiers were evaluated with stratified five-fold cross-validation under k-anonymity, Gaussian perturbation, DP-configured training at three target privacy budgets, randomised response, simulated federated learning and a hybrid configuration. The logistic-regression baseline achieved 85.49 ± 0.18% accuracy, macro-F1 0.8493 and AUC-ROC 0.9280. Utility was almost fully preserved: k-anonymity, Gaussian perturbation and the $\varepsilon = 5.0$ target each cost less than 0.4 percentage points, whereas the hybrid configuration carried the largest cost (10.37 percentage points); single-mechanism results were consistent across all four classifiers (maximum difference 0.36 percentage points). Each control was mapped to the distinct attack surface it is designed to address, enabling mechanisms to be selected by attacker boundary rather than by a single privacy ranking. Measured attack resistance and formal (ε, δ) guarantees lie outside the present scope and are not claimed, and the reported estimates await independent regeneration from archived code and fold-level outputs. The benchmark, threat map and specified attack-advantage metric provide a ready foundation for reproducible reanalysis, direct inference-attack evaluation and external validation.

1. INTRODUCTION

Machine-learning models are increasingly trained on behavioural records that describe how people earn, borrow, save and spend. The same learning pipelines can themselves become attack targets: an adversary may link released records to auxiliary data (Sweeney, 2002), query a trained model to infer whether a record was used in training (Shokri et al., 2017), or inspect the updates exchanged during distributed training (Mothukuri et al., 2021). The attack target is therefore the information-processing pipeline itself rather than a conventional software asset.

Privacy controls act on different boundaries. *k*-anonymity generalises quasi-identifiers to resist linkage (Sweeney, 2002; Torra and Navarro-Arribas, 2023); randomised response perturbs values at the point of collection (Warner, 1965; Erlingsson et al., 2014); differential privacy (DP) limits the influence of any single record on an output (Dwork and Roth, 2014), with gradient

clipping and calibrated noise as the standard route to DP model training (Abadi et al., 2016); and federated averaging keeps raw data at the clients (McMahan et al., 2017) but is not itself a privacy guarantee (Shokri and Shmatikov, 2015; Li et al., 2023; El Mestari et al., 2024). No single mechanism covers every attack surface, and each can reduce predictive utility.

Financial behavioural data are a useful test case because variables such as debt burden, emergency-fund access, spending difficulty and financial confidence are simultaneously predictive and sensitive. AI-supported financial systems have been studied for personalised investment strategies and financial-health scoring (Mali and Devmane, 2025; Mali et al., 2026), privacy-preserving classification of financial behaviour has been explored in earlier conference work (Mali and Devmane, 2026), and survey-based financial well-being data have been analysed with unsupervised methods, for example in a clustering analysis of the related 2016 National Financial Well-Being Survey (Phelps and Metzler, 2024). Three gaps remain: privacy mechanisms are often evaluated on heterogeneous datasets, which hinders comparison; financial benchmarks can become circular when the target is derived from the features used by the classifier; and mechanisms are sometimes presented as interchangeable forms of protection.

A circular benchmark is a particular concern. If a clustering algorithm creates a label and a classifier is then trained on the variables that created it, a very high score can reflect reproduction of the clustering boundary rather than independent prediction. The present study therefore constructs a seven-item financial well-being (FWB) outcome independently of the 47 predictors.

Its objectives are to (i) specify an explicit threat model and evidence hierarchy for a behavioural analytics pipeline; (ii) document eight privacy configurations on a common predictor space and an independent outcome; (iii) report utility degradation, effect sizes and cross-classifier robustness as reported estimates; and (iv) state, for every result, whether it has been independently verified. The study is a descriptive privacy–utility benchmark and threat-mapping framework and does not test a hypothesis about attack resistance. The research question is therefore: *how does a defensive privacy control change the information available to an attacker and the utility available to a behavioural classifier?*

2. MATERIALS AND METHODS

Figure 1 summarises the study architecture, from behavioural data acquisition through privacy controls and model analysis to security evidence and governance.

2.1 Study Material

The empirical foundation is the National Financial Capability Study (NFCS), using the 2021 and 2024 State-by-State survey waves, publicly available from the FINRA Investor Education Foundation (2026). The analytical sample contains 52,034 valid respondents after preprocessing. The raw survey contains 126 variables; 47 are retained as predictors after exclusion of every item used to construct the FWB outcome. Wave-specific sample counts and survey-weight handling are not reported here, and the pooled analytical sample is analysed as reported.

The study is observational and based on secondary data. No individual was re-identified and no personally identifying information was used; the data serve only to evaluate the statistical behaviour of privacy-preserving learning mechanisms.

2.2 Outcome Construction and Predictor Handling

Table 1 sets out each step in constructing the FWB outcome and the predictor matrix. The seven outcome items cover satisfaction with financial condition, financial anxiety, perceived interference of money with life goals, difficulty getting by, perceived control of finances, money remaining after monthly expenses, and confidence in achieving financial goals. The 47 predictors span emergency preparedness and liquidity, debt and credit behaviour, retirement and investment planning, health-related financial stress, objective financial literacy, self-assessed financial competence and socio-economic status. Item-level scoring, the dichotomisation threshold and the imputation settings are not reported in detail, which limits exact reproducibility (Section 3.12).

The predictor matrix excludes all seven outcome items. This separation prevents the classifier from reconstructing a label from the variables that define it, so the prediction task is behavioural predictors → independent FWB outcome rather than cluster features → cluster labels. No feature is introduced after the privacy transformation, so every mechanism is compared on a fixed predictor definition.

Natural Resources for Human Health
eISSN: 2583-1194

Table 1. Construction of the FWB Outcome and Predictor Matrix	
Step	**Procedure**
1. Item selection	Seven NFCS items combined into a composite FWB score
2. Orientation	Negatively oriented items reverse-aligned
3. Aggregation	Items aggregated into one composite score
4. Dichotomisation	Binary outcome from a predefined percentile threshold
5. Predictor exclusion	All seven outcome items removed; 47 of 126 variables kept as predictors
6. Encoding and scaling	Categorical and ordinal variables encoded; continuous variables standardised
7. Missing data	Handled by a preprocessing protocol
8. Sample	52,034 respondents, 2021 and 2024 NFCS waves

FWB, financial well-being; NFCS, National Financial Capability Study.

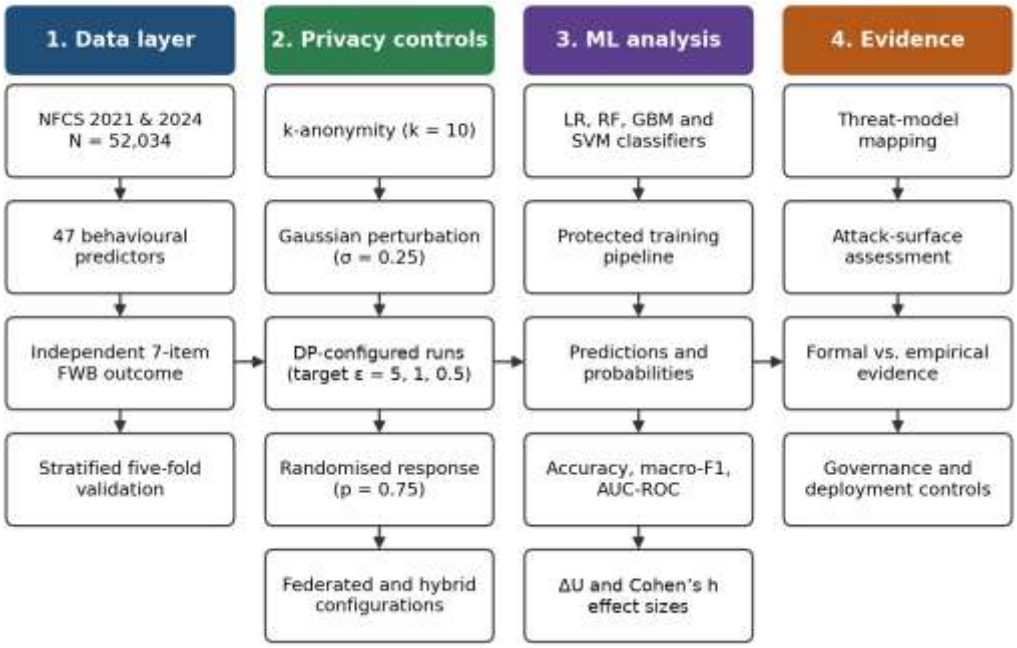


Figure 1. Threat-Aware Privacy-Preserving Architecture for Financial Well-Being (FWB) Analytics. The architecture separates the data layer, the privacy and security controls, the machine-learning analysis, and the security-evidence and governance layer. NFCS, National Financial Capability Study; LR, logistic regression; RF, random forest; GBM, gradient boosting machine; SVM, support-vector machine.

2.3 Behavioural Segmentation

K-means clustering was used only as a construct-validation step and not as the classification target. Candidate values $k = 2, \ldots, 10$ were compared using the within-cluster sum of squares, the silhouette score (Rousseeuw, 1987) and the Davies–Bouldin index (Davies and Bouldin, 1979). The selected two-cluster solution is interpreted as high-practice and low-practice behavioural

segments, and stability was assessed across repeated initialisations. The comparison between these segments and the independent FWB outcome serves as an external validity check.

2.4 Classifiers and Cross-Validation Procedure

Four supervised classifiers were evaluated: logistic regression (LR), random forest (RF) (Breiman, 2001), gradient boosting machine (GBM) (Friedman, 2001) and a radial-basis-function support-vector machine (SVM) (Cortes and Vapnik, 1995), all implemented in scikit-learn (Pedregosa et al., 2011). LR is the common privacy-evaluation baseline because it is computationally lightweight, interpretable and compatible with gradient-based privacy mechanisms. The SVM experiment uses a 15,000-sample subset per fold for computational feasibility and is therefore treated as a robustness indicator rather than a primary benchmark.

Performance was estimated by stratified five-fold cross-validation with a fixed random seed of 42. The 52,034 respondents are partitioned into five disjoint folds that each preserve the proportion of the two FWB outcome classes. In each of five iterations one fold (about 10,400 respondents) is held out for evaluation and the other four (about 41,600) are used for training; every respondent is therefore evaluated exactly once. Each reported metric is the mean ± standard deviation of the five fold-level estimates. The fold estimates share training data and are not independent, so the standard deviation describes fold-to-fold variability and understates total uncertainty. The same seed was used to generate the partition for every privacy configuration.

2.5 Threat Model

Let the protected analytics system be $\mathcal{S} = (X, Y, \mathcal{M}, T, O)$, where X is the behavioural feature matrix, Y the independent outcome, $\mathcal{M}$ the set of privacy mechanisms, T the learning (training) algorithm and O the observable system output. An attacker is represented by $\mathcal{A} = (B, V, G)$, where B is the attacker's background (prior) knowledge, V the set of observations available and G the inference goal. The security objective is not to remove all information from O but to minimise information useful for unauthorised inference while preserving enough for the authorised task, giving the empirical security–utility problem in Eq. (1):

$$\min_{M \in \mathcal{M}} \mathcal{L}_{\text{sec}}(M, \mathcal{A}) \quad \text{subject to} \quad U(M) \geq U_{\min} \tag{1}$$

where $\mathcal{L}_{\text{sec}}$ denotes an attack-loss measure when an attack is executed and $U(M)$ is the authorised predictive utility. Because no attack was executed, $\mathcal{L}_{\text{sec}}$ is not assigned a numerical value. Table 2 distinguishes four levels of evidence; the hierarchy prevents, for example, a formal DP definition from being presented as evidence that an implementation resisted a measured membership-inference attack.

Table 2. Security Evidence Levels Used in the Study

Evidence level	Meaning	Status in the present study
Formal	Security property follows from a stated mathematical definition and its implementation assumptions	DP and LDP definitions; k-anonymity equivalence-class condition
Empirical	Property is measured directly on the experimental system	Utility, AUC, macro-F1, effect sizes, cross-model robustness (reported estimates)
Threat-mapped	Mechanism is logically associated with an attacker class but the attack is not executed	Membership, attribute and gradient-reconstruction threats
Deployment-required	Additional control is required before production security claims	Secure aggregation, formal privacy accountant, attack red-team evaluation

DP, differential privacy; LDP, local differential privacy.

The threat model contains five attacker classes (A1–A5), specified in Table 3 by prior knowledge, observations and inference goal. The analyst who trains the model is assumed trusted with respect to A1 and A3; the data collector is untrusted for A2; the

aggregator and other clients are untrusted for A4 and A5. Attacks were not executed: the classes define what each control is intended to address, not what was demonstrated. Out of scope are data-poisoning and backdoor attacks, adaptive white-box attackers who combine several surfaces, side-channel attacks and deployment-level compromises such as key theft. The mechanisms are not assumed to protect against all five classes equally.

Table 3. Threat Model: Attacker Classes, Protection Boundaries and Evidence Status

Attacker class	Knowledge (*B*)	Observations (*V*)	Goal (*G*)	Control and boundary
A1 Linkage attacker	Auxiliary data with quasi-identifiers	Released (generalised) records	Re-identify respondents	*k*-anonymity ($k = 10$): protects specified QI classes only; no attribute or membership protection.
A2 Honest-but-curious collector	Survey instrument; population priors	Submitted responses	Infer sensitive answers	Randomised response ($p = 0.75$): collection boundary only; $\varepsilon_{LDP} \approx 1.099$ per binary report.
A3 Model auditor	Query access; data distribution	Predictions and confidence scores	Membership inference	DP-configured training (target ε): a guarantee holds only if the mechanism and accounting satisfy the (ε, δ)-DP definition; not established here.
A4 Curious federated participant	Own data; global model	Shared model and received updates	Infer other clients' records	Federated learning with clipping and noise: reduces raw-data exposure; update leakage untested.
A5 Insider or aggregation node	Access to client updates	Per-client gradients	Reconstruct records	Clipping and DP noise: needs a validated accountant and secure aggregation.

QI, quasi-identifiers; DP, differential privacy; LDP, local differential privacy.

2.6 Privacy Configurations

Table 4 summarises, for each configuration, the implementation and the protection boundary and guarantee status. The eight configurations are not ordered by privacy strength and must not be treated as equivalent forms of protection: they differ in the attack they address, in whether a formal definition applies, and in where in the pipeline they act. Equal utility loss therefore does not imply equal protection.

2.6.1 k-Anonymity

k-anonymity is applied to the quasi-identifiers using generalisation. The requirement on the transformed dataset D^* is given in Eq. (2), with $k = 10$:

$$\forall t \in D^*: |\{t' \in D^* : t'[\mathrm{QI}] = t[\mathrm{QI}]\}| \geq k \tag{2}$$

where $t[\mathrm{QI}]$ denotes the quasi-identifier values of record t. The mechanism addresses linkage through the specified quasi-identifiers; it does not establish protection against attribute or membership inference.

2.6.2 Gaussian Perturbation

A statistical perturbation control adds independent Gaussian noise to standardised predictor values (Eq. (3)):

$$\tilde{x} = x + \mathcal{N}(0, \sigma^2 \mathbf{I}), \qquad \sigma = 0.25 \tag{3}$$

Without a formal sensitivity calculation and privacy accounting, a fixed σ does not justify an (ε, δ)-DP claim, so this control is treated as empirical perturbation only.

Table 4. Implementation and Guarantee Status of Each Privacy Configuration

Configuration	Implementation	Boundary and guarantee status
***k*-anonymity**	QI (income bracket, age group, education) generalised so each class has at least 10 records	Linkage via QI. Formal *k*-anonymity condition; no attribute or membership protection.
Gaussian perturbation	Noise $\mathcal{N}(0, \sigma^2 I)$, $\sigma = 0.25$, added to standardised predictors	Statistical perturbation only; no DP claim.
DP-configured runs	Clipping $C = 1$; $\delta = 10^{-5}$; target $\varepsilon = 5.0, 1.0, 0.5$	Target ε values are configuration labels only; no formal (ε, δ) guarantee.
Randomised response	Binary value kept with $p = 0.75$, otherwise flipped	Collection boundary. $\varepsilon_{LDP} = \ln 3 \approx 1.099$ per binary attribute; the budget grows with the number of attributes.
Federated learning simulation	Federated averaging; 3 clients; 5 rounds; clipped, noised updates; one machine	Reduces central raw-data exposure; no formal guarantee; update leakage untested.
Hybrid	*k*-anonymity, randomised response and a DP-configured step	Layered empirical configuration; no composed guarantee.

DP, differential privacy; QI, quasi-identifiers; LDP, local differential privacy.

2.6.3 Differential Privacy

For neighbouring datasets D and D' differing by one record, a randomised mechanism M satisfies (ε, δ)-DP (Dwork and Roth, 2014) when Eq. (4) holds for all measurable output sets S:

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D') \in S] + \delta \tag{4}$$

The study reports DP-configured runs with gradient clipping $C = 1$, $\delta = 10^{-5}$ and target ε of 5.0, 1.0 and 0.5. These are configuration labels, not verified privacy budgets: the sampling scheme, number of steps, noise multiplier and composition were not tracked by a privacy accountant. No formal (ε, δ) guarantee can therefore be certified, and the ε labels must not be read as established cumulative guarantees. Runs of this kind are simulated privacy controls, distinct from mechanisms whose guarantee has been proved and accounted.

2.6.4 Randomised Response

For a binary value v, randomised response reports the true value with probability $p = 0.75$ and its complement otherwise (Eq. (5); Warner, 1965):

$$\tilde{v} = \begin{cases} v, & \text{with probability } p, \\ 1 - v, & \text{with probability } 1 - p \end{cases} \tag{5}$$

For a single binary report the corresponding local privacy parameter is given by Eq. (6) (Erlingsson et al., 2014). This control is most relevant when the collection boundary is untrusted.

$$\varepsilon_{\text{LDP}} = \ln \frac{p}{1 - p} = \ln 3 \approx 1.099 \tag{6}$$

2.6.5 Federated Learning Simulation

Federated averaging (McMahan et al., 2017) was simulated with three client partitions and five communication rounds. The aggregation rule is given in Eq. (7):

$$w^{t+1} = \sum_{i=1}^{m} \frac{n_i}{n} w_i^{t+1}, \qquad n = \sum_{i=1}^{m} n_i \tag{7}$$

where $m = 3$ is the number of clients, n_i is the number of training records at client i and w_i^{t+1} is the locally updated model of client i in round t. DP-style noise is applied to clipped local updates. The implementation is a single-machine simulation; it evaluates the computational and statistical behaviour of federated aggregation but not a geographically distributed deployment, and no cumulative ε is asserted without a formal accountant.

2.6.6 Hybrid Configuration

The hybrid configuration combines k-anonymity, randomised response and a DP-configured step. The composed privacy budget of these layers was not computed, so the configuration is reported only as a named experimental configuration, and no formal composed privacy guarantee is claimed.

2.7 Evaluation Algorithm

The evaluation procedure is given in Algorithm 1, using explicit input and output declarations followed by numbered procedural stages.

2.8 Statistical Analysis

For $K = 5$ folds, mean accuracy is defined by Eq. (8); macro-F1 and AUC-ROC are likewise reported as fold means. Utility degradation is defined by Eq. (9):

$$\mathrm{Acc} = \frac{1}{K}\sum_{k=1}^{K} \frac{\mathrm{TP}_k + \mathrm{TN}_k}{\mathrm{TP}_k + \mathrm{TN}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{8}$$

$$\Delta U = \mathrm{Acc}_{\mathrm{baseline}} - \mathrm{Acc}_{\mathrm{privacy}} \tag{9}$$

ΔU is expressed in percentage points. Paired McNemar testing (McNemar, 1947) was specified in the analysis plan, with a Bonferroni-adjusted threshold $\alpha_c = 0.05/8 = 0.00625$ for the eight protected configurations. The paired predictions and test outputs are not available, so no McNemar p-values, significance claims or inferential conclusions are reported. Effect size is measured by Cohen's h (Eq. (10); Cohen, 1988), computed from the rounded accuracies reported in this paper:

$$h = 2\arcsin\sqrt{p_1} - 2\arcsin\sqrt{p_2}. \tag{10}$$

Because Cohen's conventional benchmarks (0.2, 0.5 and 0.8) are coarse for accuracy differences of this size, descriptive tiers specific to this study were used: negligible ($h < 0.02$), small ($0.02 \le h < 0.08$), small–medium ($0.08 \le h < 0.20$) and medium ($h \ge 0.20$). These tiers are operational conventions of this study and not Cohen's thresholds; under Cohen's own conventions only the hybrid configuration ($h = 0.263$) would reach the "small" benchmark of 0.2, and no configuration would reach the "medium" benchmark of 0.5.

Natural Resources for Human Health
eISSN: 2583-1194

Algorithm 1. Threat-aware privacy control and security evaluation.

Input: dataset D, predictors X, independent outcome Y, mechanisms $\mathcal{M}$, classifiers $\mathcal{C}$

Output: security–utility profile $\mathcal{P}$ and mechanism-to-threat mapping

Stage 1 – Dataset preparation

Remove all variables used to construct Y from X
Encode, impute and standardise the fixed predictor matrix
Define quasi-identifiers QI = {income, age, education}
Partition D using stratified five-fold cross-validation

Stage 2 – Baseline model

Train each classifier $c \in \mathcal{C}$ on the unprotected training folds
Select LR as the common privacy-evaluation baseline
Store baseline predictions, accuracy, macro-F1 and AUC-ROC

Stage 3 – Apply privacy control

for each mechanism $m \in \mathcal{M}$ **do**
    Apply m to the training pipeline, keeping the predictor set and the outcome definition fixed
    k-anonymity: generalise QI and enforce $|QI\text{-class}| \geq 10$
    Gaussian: add $\mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.25$
    DP: apply clipping $C = 1$ with configured ε, δ
    Randomised response: report the true binary value with $p = 0.75$
    Federated learning: aggregate three simulated clients for five rounds
    Hybrid: apply the specified layered transformations
end for

Stage 4 – Utility evaluation and threat mapping

for each protected configuration **do**
    Compute accuracy, macro-F1, AUC-ROC and ΔU
    Compute Cohen's h (paired McNemar testing specified but not reported)
    Map the mechanism to the attacker classes it is intended to mitigate
    Store the resulting record in $\mathcal{P}$
end for
Repeat the mechanism comparison for RF, GBM and SVM
Report the maximum inter-model variation and the limitations of the threat claims
return $\mathcal{P}$

For membership inference, the attack advantage is defined in Eq. (11), where TPR_{MI} and FPR_{MI} are the true- and false-positive rates of the attack at a specified operating point:

$$Adv_{MI} = TPR_{MI} - FPR_{MI} \tag{11}$$

A privacy control would be considered empirically useful when the attack advantage decreases without an unacceptable reduction in authorised predictive utility. This quantity is not reported because no membership-inference experiment was executed; Eq. (11) specifies the metric to be reported when it is.

2.9 Evidence Status of Reported Results

To separate reported performance estimates from independently verified results, every numerical result in this paper carries one of the statuses defined in Table 5. "Reported" means the value comes from the authors' original analysis and has not been independently regenerated from archived fold-level outputs. Reported values should be treated as provisional utility estimates and not as verified evidence of privacy protection.

Table 5. Evidence Status of the Results Reported in This Paper

Result	Basis	Status
Accuracy, macro-F1 and AUC-ROC for the baseline and the eight configurations (Sections 3.2 and 3.4)	Original analysis; mean ± SD over five folds	Reported; not independently regenerated
Cross-classifier utility losses (Section 3.8)	Original analysis	Reported; hybrid evaluated with LR only
ΔU	Difference of reported means	Recomputed from rounded accuracies; agrees within ±0.01 pp
Cohen's h and effect tiers	Cohen's h applied to reported accuracies	Recomputed from rounded accuracies; not from unrounded fold predictions
$\varepsilon_{LDP} \approx 1.099$; $a_c = 0.00625$	Analytic	Analytically verified
ROC curves and confusion matrix	Fold-level predictions not retained (Section 3.5)	Not shown; AUC values reported in Section 3.4
Cluster difference (66.7% vs 9.8%, $p < 0.001$)	Original analysis	Reported; not independently regenerated
McNemar tests	Specified in the analysis plan	Not reported; no inferential claims
Empirical attack resistance	No attack executed	Not evaluated
Formal (ε, δ) guarantee	No privacy accounting performed	Not established

SD, standard deviation; ROC, receiver operating characteristic; pp, percentage points.

2.10 Reproducibility Controls

The study reports Python 3.10, scikit-learn 1.3, NumPy 1.24 and SciPy 1.11 with random seed 42 (Table 6). The analysis code is available from the corresponding author on reasonable request. Full independent verification requires archiving the exact preprocessing pipeline, predictor list, outcome construction, model parameters, privacy transformations, fold assignments and fold-level outputs, and regenerating every table and figure from them.

Table 6. Reproducibility Controls for the Benchmark

Control	Specification
Dataset	NFCS 2021 and 2024 waves; final analytical N = 52,034
Predictors	47 variables; all seven FWB outcome items excluded
Validation	Stratified five-fold cross-validation; seed 42
Software	Python 3.10; scikit-learn 1.3; NumPy 1.24; SciPy 1.11
LR	C = 1.0 (scikit-learn inverse regularisation strength), LBFGS solver, maximum 500 iterations
RF	100 estimators, maximum depth 10
GBM	50 estimators, learning rate 0.1, maximum depth 3
SVM	RBF kernel, C = 1.0, 15,000-sample subset per fold
Privacy configurations	k = 10; σ = 0.25; target ε = 5.0, 1.0, 0.5 ($\delta = 10^{-5}$, C = 1); p = 0.75; FL simulation (3 clients, 5 rounds); hybrid
Statistics	Cohen's h; Bonferroni-adjusted McNemar threshold specified but not reported

FWB, financial well-being; FL, federated learning; LR, logistic regression; RF, random forest; GBM, gradient boosting machine; SVM, support-vector machine.

3. RESULTS AND DISCUSSION

3.1 Behavioural Segmentation

The two-cluster solution was selected from candidate values k = 2 to 10. The silhouette score was 0.122 and the Davies–Bouldin index 3.031, indicating weak-to-moderate separation in the behavioural feature space; the selected partition remained stable across repeated initialisations. External validation against the independent FWB outcome showed a 56.9-percentage-point difference in the high-FWB rate between the high-practice and low-practice clusters (66.7% versus 9.8%; $p < 0.001$). **Figure 2** summarises the clustering diagnostics and external outcome separation. The clustering is not used as the supervised target; its purpose is to show that the behavioural feature space carries structure that is externally associated with financial well-being.

3.2 Baseline Classification

Table 7 shows the reported baselines for the four classifiers. All models lie within a narrow accuracy range (84.91–85.71%), which supports the robustness of the underlying behavioural signal. Logistic regression achieved 85.49 ± 0.18% accuracy, macro-F1 0.8493 and AUC-ROC 0.9280 and was retained as the common privacy-evaluation baseline. Using a fixed, independently defined outcome reduces the risk that a privacy mechanism appears effective merely because it preserves a circular classification boundary.

3.3 Behavioural Feature Importance

Permutation importance identified difficulty paying monthly bills as the strongest predictor (PI = 0.0607 ± 0.0026), followed by perceived debt burden (PI = 0.0275 ± 0.0024) and access to emergency funds (PI = 0.0124 ± 0.0014). **Figure 3** shows the top predictors and the importance summed by feature domain. The result has a security implication: high-value behavioural attributes are also high-value privacy targets, and a mechanism that protects only demographic quasi-identifiers may leave the most informative attributes unchanged.

3.4 Privacy–Utility Evaluation

Table 8 gives the reported quantitative benchmark for the eight configurations summarised by mechanism in Table 4 (three of the eight are DP targets). *k*-anonymity preserved baseline utility to rounding precision; Gaussian perturbation and the $\varepsilon = 5.0$ target produced negligible accuracy changes; stronger configurations produced progressively larger utility costs. **Figure 4** visualises the accuracy gradient. These results compare utility only: they do not rank the configurations by privacy strength.

Small numerical differences, such as those produced by Gaussian noise and the $\varepsilon = 5.0$ target, could be statistically detectable in a sample of 52,034 respondents while remaining practically negligible. Because McNemar results are not reported, this paper draws no inferential conclusion and relies on effect sizes and descriptive utility change only.

3.5 Multi-Metric Consistency

Figure 5 displays accuracy, macro-F1 and AUC-ROC across configurations. These metrics were not regenerated from common saved fold-level outputs, so their alignment is descriptive and must be verified by rerunning the analysis. ROC curves and confusion matrices are not shown because fold-level predictions were not retained; AUC values are reported in Table 8.

Table 7. Reported Baseline Classifier Performance under Stratified Five-Fold Cross-Validation (Mean ± SD)

Model	Accuracy (%)	Macro-F1	AUC-ROC
Logistic regression (LR; baseline)	85.49 ± 0.18	0.8493 ± 0.002	0.9280 ± 0.001
Random forest (RF)	85.71 ± 0.28	0.8500 ± 0.002	0.9305 ± 0.001
Gradient boosting machine (GBM)	85.63 ± 0.23	0.8497 ± 0.002	0.9303 ± 0.001
Support-vector machine (SVM)[a]	84.91 ± 0.38	0.8441 ± 0.004	0.9228 ± 0.002

[a] Trained on a 15,000-sample subset per fold; treated as a robustness indicator. Values are reported estimates (Table 5).

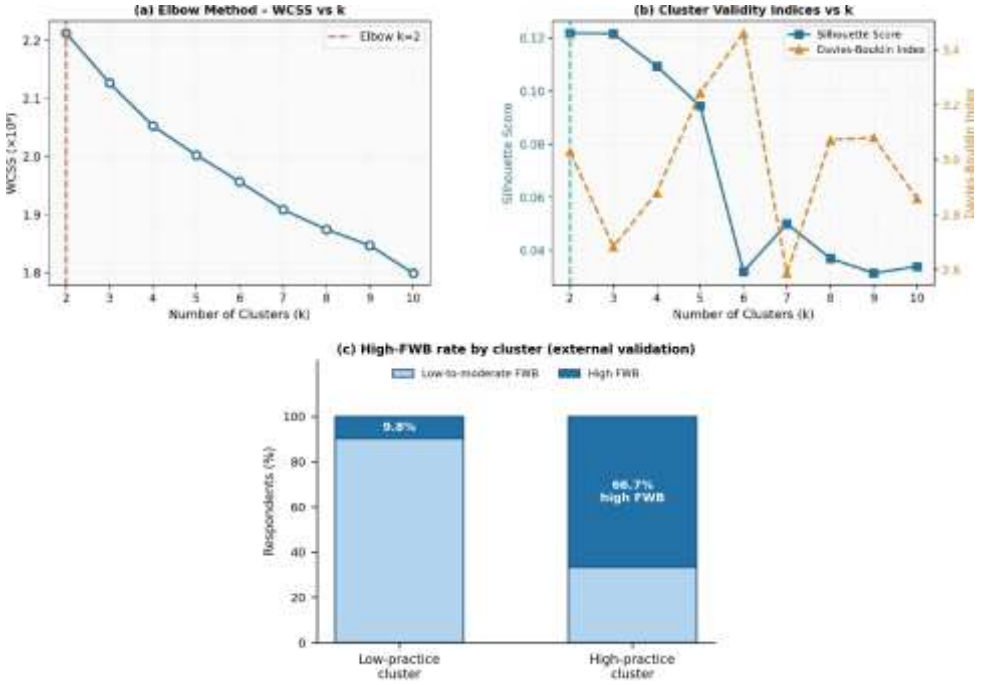


Figure 2. Behavioural Clustering Validation. (a) Elbow analysis; (b) cluster validity indices (silhouette score and Davies–Bouldin index); (c) independent high-FWB rate in each of the two clusters.

Natural Resources for Human Health
eISSN: 2583-1194

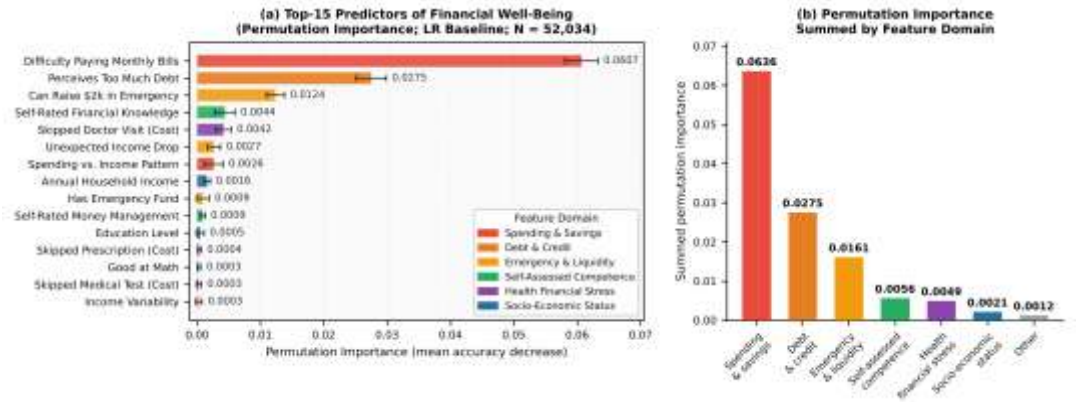


Figure 3. Permutation Importance for the Independent FWB Prediction Task (LR Baseline). (a) The 15 most important predictors; (b) importance summed by feature domain. Spending and savings behaviour and debt and credit behaviour dominate the predictive signal.

Table 8. Reported Privacy–Utility Evaluation (Mean ± SD over the Five-Fold Protocol; LR Baseline)

Configuration	Accuracy (%)	Macro-F1	AUC-ROC	ΔU (pp)	Cohen's h	Effect tier
Baseline (no privacy)	85.49 ± 0.18	0.8493 ± 0.002	0.9280 ± 0.001	—	—	—
k-Anonymity (k = 10)	85.50 ± 0.19	0.8494 ± 0.002	0.9280 ± 0.001	0.00	0.000	Negligible
Gaussian (σ = 0.25)	85.17 ± 0.14	0.8458 ± 0.001	0.9243 ± 0.001	0.33	0.009	Negligible
DP-configured (target ε = 5.0)	85.26 ± 0.13	0.8469 ± 0.001	0.9256 ± 0.001	0.23	0.007	Negligible
FL + DP simulation	83.99 ± 0.18	0.8373 ± 0.002	0.9205 ± 0.001	1.50	0.042	Small
DP-configured (target ε = 1.0)	82.39 ± 0.36	0.8165 ± 0.004	0.8975 ± 0.002	3.10	0.085	Small–medium
RR (p = 0.75)	80.81 ± 0.25	0.8006 ± 0.003	0.8816 ± 0.002	4.68	0.125	Small–medium
DP-configured (target ε = 0.5)	77.95 ± 0.20	0.7681 ± 0.003	0.8506 ± 0.002	7.54	0.196	Small–medium
Hybrid	75.12 ± 0.41	0.7374 ± 0.004	0.8186 ± 0.003	10.37	0.263	Medium

$\Delta U = \text{Acc}_{\text{baseline}} - \text{Acc}_{\text{privacy}}$ in percentage points (pp), reported from unrounded means and therefore possibly ±0.01 from the difference of the rounded accuracies (Gaussian: 85.49 − 85.17 = 0.32). Cohen's h was recomputed from the rounded accuracies and tiers follow Section 2.8. DP-configured runs carry the stated target ε only; no privacy accounting was performed. All values are reported estimates (Table 5).

3.6 Single-Layer versus Layered Configurations

The hybrid configuration has the largest reported accuracy loss (10.37 percentage points; h = 0.263), against 4.68 points (h = 0.125) for randomised response alone and 0.33 points (h = 0.009) for Gaussian perturbation alone (Table 8). This is a descriptive comparison and not a factorial ablation: pairwise combinations were not run, the component order was not varied, and no

causal attribution to individual layers is made. Nor does it show that the hybrid is more secure, because no common attack metric was measured across layers.

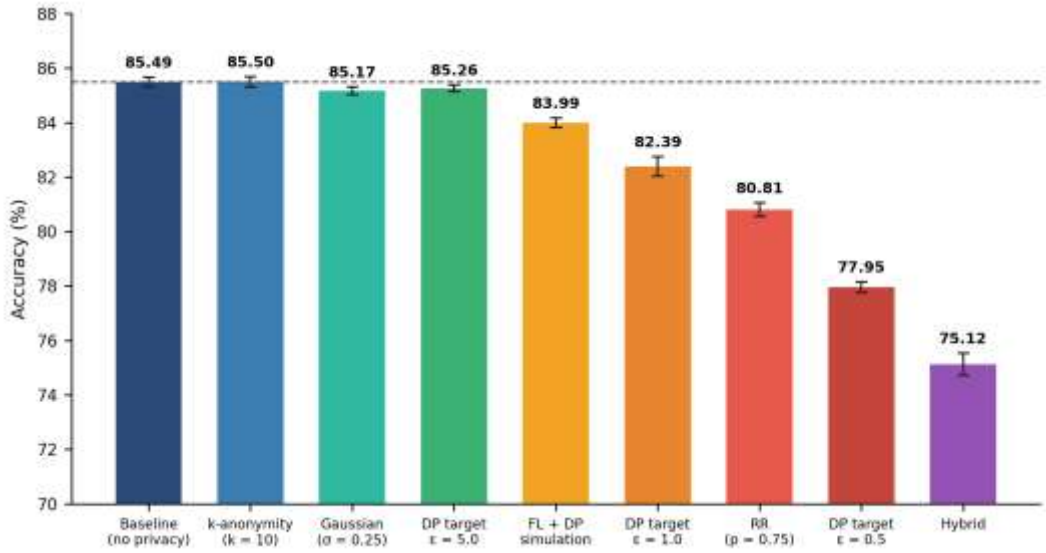


Figure 4. Reported Privacy–Utility Trade-Off across the Evaluated Configurations. LR baseline, stratified five-fold cross-validation, seed 42. Bars show mean accuracy, error bars ±1 SD across folds, and the dashed line the non-private baseline (85.49%). The y-axis starts at 70%.

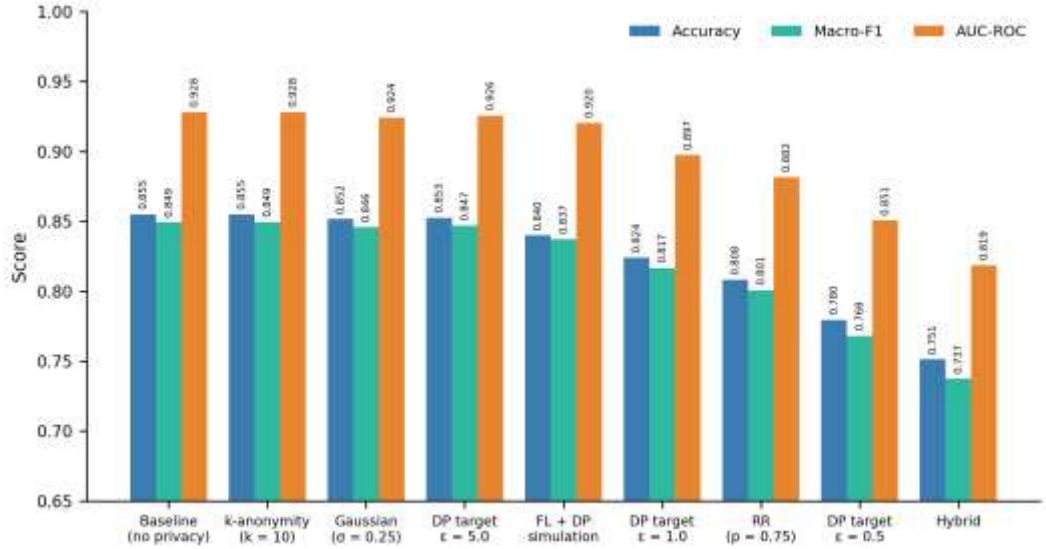


Figure 5. Reported Accuracy, Macro-F1 and AUC-ROC across the Privacy Configurations (LR Baseline). The three metrics follow the same ordering. The y-axis starts at 0.65.

3.7 Effect-Size Profile

Figure 6 displays Cohen's *h* recalculated from the rounded accuracies in Table 8. Because unrounded fold predictions are unavailable, these are rounded-value recalculations and not independently verified effect-size estimates. Under the study-specific tiers, the $\varepsilon = 0.5$ target ($h = 0.196$) falls just below the 0.20 boundary of the medium tier.

3.8 Cross-Classifier Robustness

Across the seven single-mechanism configurations, the utility losses are broadly similar for LR, RF, GBM and SVM (Table 9); the largest difference between classifiers is 0.36 percentage points (randomised response). The hybrid configuration was evaluated with the LR baseline only (Table 8), so its cross-classifier stability is not assessed here.

Table 9. Reported Cross-Classifier Utility Degradation (Percentage Points) Relative to Each Model's Non-Private Baseline

Mechanism	LR	RF	GBM	SVM	Max. difference
k-Anonymity	0.00	0.00	0.00	0.00	0.00
Gaussian	0.33	0.35	0.32	0.37	0.05
DP-configured (ε = 5.0)	0.23	0.25	0.22	0.28	0.06
FL + DP	1.50	1.55	1.48	1.67	0.19
DP-configured (ε = 1.0)	3.10	3.19	3.07	3.35	0.28
RR (p = 0.75)	4.68	4.75	4.62	4.98	0.36
DP-configured (ε = 0.5)	7.54	7.68	7.49	7.84	0.35

Max. difference = largest minus smallest value across LR, RF, GBM and SVM. Values are reported estimates (Table 5).

3.9 Threat-to-Control Mapping

The threat model in Table 3 and the attack-surface relationship in **Figure 7** link each attacker class to its observable data, its intended control and the evidence needed to support a security claim. Only the utility consequences of the controls were measured; no row of the mapping has been tested by an attack, and each control is mapped to the boundary it is intended to protect and not to a demonstrated defence.

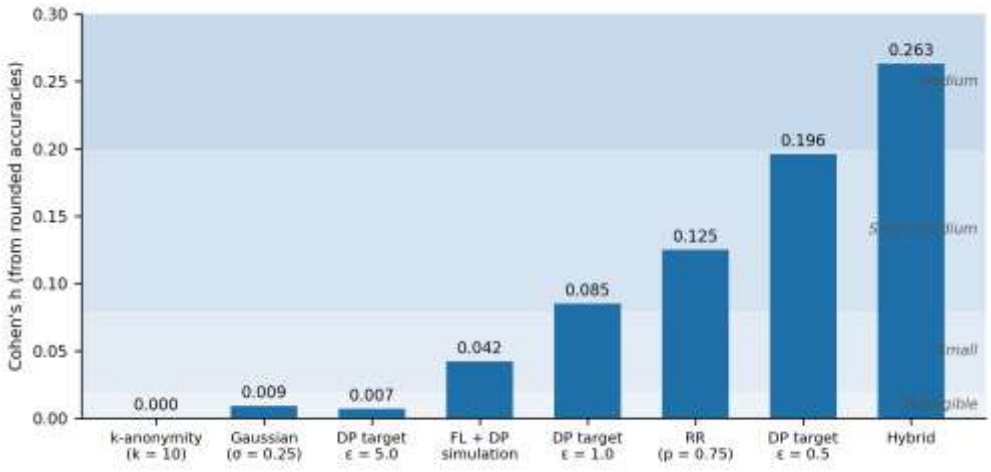


Figure 6. Effect-Size Profile of Privacy-Induced Utility Change Relative to the Non-Private Baseline. Cohen's h was recalculated from rounded accuracies; shaded bands indicate the study-specific tiers defined in Section 2.8.

3.10 Security Governance and Deployment Considerations

The empirical effect sizes provide engineering evidence about utility loss, not a legal compliance certificate. In a production financial service, the selected privacy mechanism should be combined with access control, retention limits, encryption in transit

and at rest, audit logging, key management, incident response, secure model deployment and documented data-governance procedures. The mechanism should be selected from the attacker's observation boundary rather than from a generic privacy ranking. Table 10 summarises the deployment controls and their intended protection boundary.

Table 10. Deployment Controls That Complement the Evaluated Privacy Mechanisms

Security layer	Required deployment control
Data access	Least-privilege access, role separation and auditable data queries
Data transport	Authenticated and encrypted channels; certificate and key lifecycle management
Data storage	Encryption at rest, retention limits and controlled backup access
Model interface	Rate limiting, output minimisation and monitoring for repeated inference queries
Federated learning	Secure aggregation, client authentication, update clipping and formal privacy accounting
Privacy accounting	Mechanism-specific sensitivity analysis, composition accounting and versioned privacy parameters
Security testing	Membership-inference, attribute-inference and update-leakage red-team tests before deployment

3.11 Discussion

Why the non-circular benchmark matters. A circular benchmark can produce apparently exceptional accuracy while measuring only how well a classifier reproduces the clustering algorithm; such a result is weak evidence for a security system because the target does not represent an independently observable state. The present design instead evaluates whether behavioural information predicts an external outcome while privacy mechanisms progressively restrict the information available to the classifier.

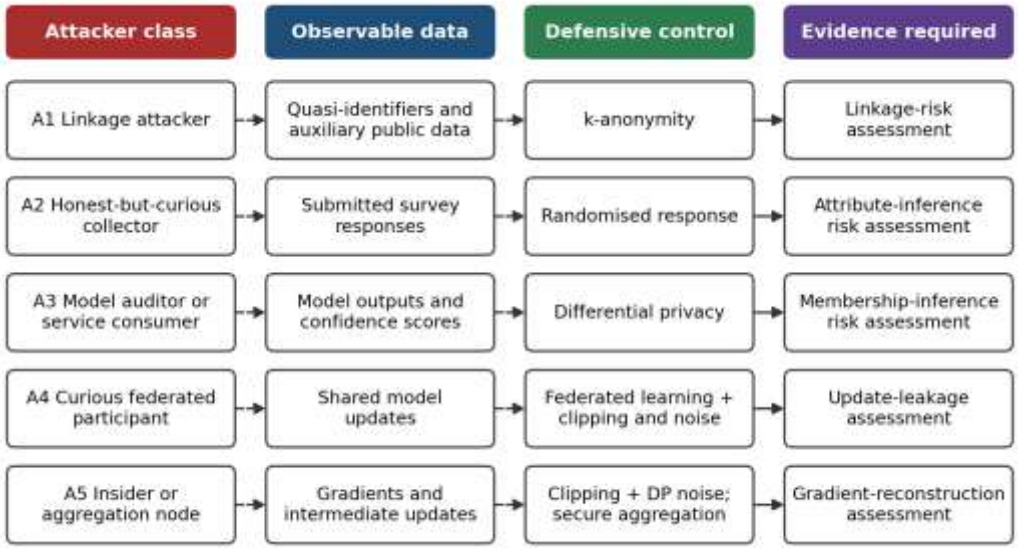


Figure 7. Attack Surface and Security-Evidence Mapping across the Behavioural Analytics Pipeline. The figure links each attacker class (A1–A5) to the observable data, the defensive control and the evidence required to support a security claim.

Interpretation of the privacy gradient. The reported values suggest three descriptive utility-cost regions: near-zero loss for k-anonymity, Gaussian perturbation and the $\varepsilon = 5.0$ target; intermediate losses for the FL + DP simulation and the $\varepsilon = 1.0$

target; and larger losses for randomised response, the $\varepsilon = 0.5$ target and the hybrid configuration. These groupings describe reported utility only. They do not rank verified privacy strength, because the configurations protect different boundaries (Table 4) and the DP implementation and accounting have not been established.

Why k-anonymity preserves utility here. The near-zero loss is dataset-specific. The transformation targets age group, income bracket and education level, whereas the strongest predictive signal comes from behavioural variables such as difficulty paying monthly bills and perceived debt burden. This should not be read as evidence that *k*-anonymity is universally utility-preserving.

Federated learning versus local randomisation. The federated simulation lost less accuracy than randomised response. A plausible explanation is that local randomisation corrupts input values before learning, whereas federated aggregation lets each local model learn from unperturbed local observations before update aggregation. This is a hypothesis, not a tested mechanism, and it is not proof that federated learning is intrinsically more private; the two mechanisms should be compared under matched attacker assumptions.

Hybrid protection and composition. The hybrid configuration produced the largest utility reduction, as expected when several transformations remove complementary information. Without a formal composition analysis, the 10.37-percentage-point loss is treated as an empirical result, no composed privacy budget is claimed, and the loss is not described as multiplicative noise amplification.

Behavioural predictors as security-sensitive assets. The most useful predictors are not conventional identifiers. Difficulty paying bills and perceived debt burden are behavioural states that may be highly sensitive even if they cannot uniquely identify an individual, so a system that anonymises age and income but exposes detailed behavioural variables may remain vulnerable to sensitive-attribute inference.

Practical implications. A defender should first identify the observation boundary available to the attacker and then select the mechanism that directly reduces that observation: generalisation for linkage, local randomisation for an untrusted collector, differential privacy for formal bounding of record-level influence, and federated learning where raw data should remain distributed. Layering can reduce several exposure paths but consumes utility. These statements describe design intent under the threat model, not demonstrated protection.

Relationship to prior conference work. An earlier conference paper examined privacy-preserving financial behaviour classification (Mali and Devmane, 2026). The present study extends it with a non-circular independent outcome design, explicit threat modelling, cross-classifier analysis, effect-size reporting and a security-boundary interpretation of each mechanism.

3.12 Limitations

The NFCS represents the United States only, federated learning was simulated on a single machine, and the SVM used subsampling. No inference attack was executed, so the study identifies protection boundaries and not empirical attack resistance, and the DP-configured, Gaussian and hybrid runs have no privacy accountant, so no (ε, δ) guarantee is established. The utility-loss estimates come from one survey and one seed-controlled partition without repeated-seed confidence intervals; fold standard deviations understate total uncertainty; item-level outcome scoring, class prevalence and fold-wise preprocessing are not reported in detail (Table 1); Cohen's *h* was computed from rounded accuracies; the hybrid configuration was evaluated with LR only; and McNemar results and ROC curves are not reported. Utility loss is not a measure of privacy gained.

4. CONCLUSION

This study presents a threat-mapped privacy–utility benchmark for financial-behaviour analytics built on an independently constructed FWB outcome. On 52,034 NFCS respondents, the reported estimates show a descriptive utility-cost gradient across eight configurations: negligible loss for *k*-anonymity, Gaussian perturbation and the $\varepsilon = 5.0$ target; intermediate loss for the FL + DP simulation and the $\varepsilon = 1.0$ target; and larger losses for randomised response, the $\varepsilon = 0.5$ target and the hybrid configuration (10.37 percentage points).

These are provisional estimates: they have not been independently regenerated from archived code and fold-level outputs. The configurations protect different boundaries and are not equivalent forms of protection. No inference attack was executed, no empirical attack resistance is claimed, and no formal (ε, δ) guarantee is established because privacy accounting was not performed. The paper should be read as a preliminary benchmark and threat-mapping framework, not as a validated security

certification. Conclusions about practical privacy protection require, first, archived code, fold indices and fold-level outputs from which every table and figure can be regenerated; second, direct evaluation of inference attacks (shadow-model membership inference, attribute inference with a hidden sensitive attribute, and federated-update reconstruction), reporting attack advantage (Eq. (11)) with repeated seeds and confidence intervals; third, formal privacy accounting for all DP-configured and composed mechanisms; and fourth, external validation, including non-US datasets and distributed federated experiments with non-IID clients.

ACKNOWLEDGEMENTS

The authors acknowledge the FINRA Investor Education Foundation for making the National Financial Capability Study data publicly available. No external funding was received for this study.

AUTHOR CONTRIBUTIONS

Nilesh Mali: Conceptualization, Data Curation, Formal Analysis, Investigation, Software, Validation, Visualization, Writing—Original Draft.

Vidyullata Devmane: Supervision, Methodology, Writing—Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

This study used publicly available, de-identified secondary NFCS data. No participant recruitment, direct contact, intervention or access to identifying information occurred; ethical approval and informed consent were therefore not required.

DATA AVAILABILITY STATEMENT

The 2021 and 2024 NFCS State-by-State datasets, questionnaires and methodological documentation are publicly available from the FINRA Investor Education Foundation (https://www.finrafoundation.org/data-and-downloads; accessed 30 March 2026).

REFERENCES

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L., 2016. Deep Learning with Differential Privacy. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. https://doi.org/10.1145/2976749.2978318

[2] Breiman, L., 2001. Random Forests. *Machine Learning* 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[3] Cohen, J., 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, NJ.

[4] Cortes, C., Vapnik, V., 1995. Support-Vector Networks. *Machine Learning* 20(3), 273–297. https://doi.org/10.1007/BF00994018

[5] Davies, D.L., Bouldin, D.W., 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[6] Dwork, C., Roth, A., 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* 9(3–4), 211–407. https://doi.org/10.1561/0400000042

[7] El Mestari, S.Z., Lenzini, G., Demirci, H., 2024. Preserving Data Privacy in Machine Learning Systems. *Computers & Security* 137, 103605. https://doi.org/10.1016/j.cose.2023.103605

[8] Erlingsson, Ú., Pihur, V., Korolova, A., 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054–1067. https://doi.org/10.1145/2660267.2660348

[9] FINRA Investor Education Foundation, 2026. National Financial Capability Study: Data and Downloads. https://www.finrafoundation.org/data-and-downloads (accessed 30 March 2026).

[10] Friedman, J.H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[11] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., He, B., 2023. A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. *IEEE Transactions on Knowledge and Data Engineering* 35(4), 3347–3366. https://doi.org/10.1109/TKDE.2021.3124599

[12] Mali, N., Devmane, V., 2025. A Systematic Exploration of AI-Powered Personalized Investment Strategies. In: *Proceedings of the 2025 3rd International Conference on Inventive Computing and Informatics (ICICI)*, pp. 404–409. https://doi.org/10.1109/ICICI65870.2025.11069526

[13] Mali, N., Devmane, V., 2026. Evaluating Privacy-Preserving Techniques for Secure Financial Behavior Classification. In: *Proceedings of the 2026 4th International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pp. 1015–1020. https://doi.org/10.1109/ICSSAS68835.2026.11559344

[14] Mali, N., Patel, S., Khandare, V., Dhavde, Y., Bhuvad, P., 2026. FinSight: Personal Financial Health Scoring and Advisory System Using Machine Learning. In: *Proceedings of the 2026 7th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 1942–1949. https://doi.org/10.1109/ICICV68925.2026.11554902

[15] McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Agüera y Arcas, B., 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 54*, pp. 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html

[16] McNemar, Q., 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika* 12(2), 153–157. https://doi.org/10.1007/BF02295996

[17] Mothukuri, V., Parizi, R.M., Pouriyeh, S., Huang, Y., Dehghantanha, A., Srivastava, G., 2021. A Survey on Security and Privacy of Federated Learning. *Future Generation Computer Systems* 115, 619–640. https://doi.org/10.1016/j.future.2020.10.007

[18] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[19] Phelps, N., Metzler, A., 2024. An Exploratory Clustering Analysis of the 2016 National Financial Well-Being Survey. *PLOS ONE* 19(9), e0309260. https://doi.org/10.1371/journal.pone.0309260

[20] Rousseeuw, P.J., 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[21] Shokri, R., Shmatikov, V., 2015. Privacy-Preserving Deep Learning. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1310–1321. https://doi.org/10.1145/2810103.2813687

[22] Shokri, R., Stronati, M., Song, C., Shmatikov, V., 2017. Membership Inference Attacks Against Machine Learning Models. In: *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. https://doi.org/10.1109/SP.2017.41

[23] Sweeney, L., 2002. *k*-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5), 557–570. https://doi.org/10.1142/S0218488502001648

[24] Torra, V., Navarro-Arribas, G., 2023. Attribute Disclosure Risk for *k*-Anonymity: The Case of Numerical Data. *International Journal of Information Security* 22(6), 2015–2024. https://doi.org/10.1007/s10207-023-00730-x

[25] Warner, S.L., 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association* 60(309), 63–69. https://doi.org/10.1080/01621459.1965.10480775