

PRIVACY-PRESERVING EXPLAINABLE DEEP LEARNING FRAMEWORK FOR LOAN DEFAULT RISK PREDICTION

Muhammad Atif Sultan^{*1}, Shaheer Arif Nazeer², Ali Husnain³, Rashid Mehmood Gondal⁴,
Abdul Hanan Mushtaq⁵, Anam Zahoor⁶, Sahrish Khan⁷, Aqsa Ejaz⁸, Muhammad Saqib⁹,
Faizan Saleem¹⁰, Yasir Ghayor Malik¹¹

^{*1,2,5,6,7,8,9,10,11} Faculty of Information Technology, The University of Lahore, Sargodha Campus, Sargodha 40100,
Pakistan

^{3,4}University of Sargodha, Sargodha 40100, Pakistan

¹atifbhatti865@gmail.com, ²cshaheer993@gmail.com, ³alihanain90@yahoo.com,
⁴rashidmehmoodit@gmail.com, ⁵abdulhannan5556h@gmail.com, ⁶malikanam06@gmail.com,
⁷sahrishkhan1215@gmail.com, ⁸zulqarnainaqs3@gmail.com, ⁹saqibaliriaz888@gmail.com,
¹⁰faizan.saleem@cs.uol.edu.pk, ¹¹yasir.ghayor@cs.uol.edu.pk

DOI: <https://doi.org/10.5281/zenodo.23315840>

Keywords

Differential privacy; DP-SGD; explainable deep learning; loan default prediction; privacy-utility trade-off.

Article History

Received: 25 June 2026

Accepted: 16 September 2026

Published: 02 October 2026

Copyright @Author

Corresponding Author: *

Muhammad Atif Sultan

Abstract

Automated loan-approval systems increasingly rely on deep neural networks that can capture non-linear relationships between an applicant's financial history and their likelihood of default. Two obligations compete inside such a system: the model must be accurate enough to be economically useful, and it must not leak information about the individual applicants whose records it was trained on. Differential privacy (DP) is the standard mathematical language for the second obligation, but in the bank-loan-prediction literature it has almost always been bolted onto the outside of a classical ensemble model noise is added to features or to a tree ensemble's outputs rather than integrated into the training procedure of a deep network itself. This study closes that gap by training a multilayer perceptron (MLP) for loan default prediction with DP-SGD, the differentially private stochastic-gradient-descent procedure in which every individual applicant's gradient contribution is clipped and Gaussian noise is added before the network's weights are updated, so that the privacy guarantee is a property of the learning algorithm rather than of a pre-processing step. On the standard 614-record public loan-approval benchmark, a non-private baseline MLP reaches 83.3% accuracy and a ROCAUC of 0.775 (mean of five seeds), statistically indistinguishable from a Random Forest (85.0%) and a histogram gradient-boosting model (81.3%) trained on the identical split. Training the same architecture with DP-SGD across a swept noise level traces a smooth, monotonically increasing privacy-utility curve: accuracy rises from 48.5% at a strict privacy budget of $\epsilon = 0.39$ to 82.0% at a more permissive $\epsilon = 7.27$, recovering 98% of the non-private model's accuracy. Because the advanced-composition bound used for that headline sweep is deliberately conservative, a second, Rényi-DP accountant is computed for the same trained networks and reports privacy

budgets four to five times tighter for identical noise levels, and the qualitative collapse-then-recover pattern replicates on an independent 572-record external validation dataset. Three complementary explainability techniques permutation importance, integrated gradients and SHAP are applied directly to the trained network rather than to a separate surrogate tree model, and all three independently identify credit history as the dominant predictor. A secondary experiment compares three ways of handling the dataset's 2.2:1 class imbalance and finds that synthetic oversampling and class-weighting trade a small amount of aggregate accuracy for a modest improvement in the minority-class recall. The results suggest that DP-SGD is a practical, feature-agnostic route to privacy-preserving deep learning for credit-risk prediction.

INTRODUCTION

Lenders determine loan suitability by analyzing risk factors such as wage, credit background, debts, professional status. Modern software programs solve problems like this well because multi layer analytical processes detect correlations between monetary details which older manual grading systems miss. Nonetheless, these same sources help forecast repayments while exposing people to risks like personal identification theft, prejudice, financial destruction. Such details encompass earnings, union status, dependants, banking history. Current legal systems like GDPR mandate aggressive guardianship over personal information dossiers instead of relying upon basic oversight processes to ensure consumer safety from predatory practices. Organizations handle significant pressure keeping sensitive consumer profiles secure from digital exploitation or data exposure. Responsible institutions prioritize defensive strategies versus loose maintenance standards regarding confidential archives since information theft leads towards permanent fiscal disaster for unsuspecting clients caught within compromised systems today.

Differential privacy (DP) is the most widely accepted mathematical framework for that protection: it bounds, in a precise and composable way, how much any single training record can influence a model's output. Within the loan-prediction literature, DP has typically been applied at the feature level adding calibrated noise to income or loan-amount columns before a classical model such as Random Forest or CatBoost is

trained or at the output level, perturbing a model's released statistics. A 2025 benchmark study that combined Laplace- and Gaussian-mechanism DP with five ensemble classifiers on a public loan-approval dataset illustrates both the promise and the difficulty of this approach: the strongest Laplace-mechanism configuration lost more than twenty accuracy points relative to a non-private baseline, while the gentler Gaussian mechanism, paired with CatBoost, retained a substantially larger share of the non-private accuracy (Hossain et al., 2025). That contrast is informative, but it leaves a structural question unaddressed in both configurations, privacy is enforced by perturbing data or outputs that sit outside the model, rather than by constraining what the learning algorithm itself is allowed to memorise.

Deep learning offers a different, arguably more natural place to enforce that constraint: the gradient update. DP-SGD, introduced by Abadi et al. (2016), clips each training example's contribution to a mini-batch gradient to a fixed norm and adds calibrated Gaussian noise to the clipped sum before the weights are updated. Because every example is treated identically regardless of which features it touches, DP-SGD does not require a separate sensitivity analysis for income, credit history, and loan amount the way a feature-level mechanism does the privacy guarantee is a property of the optimiser, not of the input pipeline. Despite being the default mechanism for private deep learning in the wider machine-learning literature, DP-SGD has not, to the best of our knowledge, been evaluated end-to-

end on this benchmark loan-approval dataset with a tuned neural architecture and a full privacy-utility sweep.

This study fills that gap. A compact multilayer perceptron is trained for loan default prediction under an explicit, auditable DP-SGD procedure, and its accuracy is traced across a swept privacy budget from a strict $\epsilon \approx 0.4$ to a permissive $\epsilon \approx 7$. Three further design choices distinguish the study from prior feature-level or non-private treatments of this dataset: (i) all monetary fields are engineered into ratio and log-transformed features before training, so that the network sees a numerically well-conditioned input space rather than raw currency values; (ii) three strategies for handling the dataset's class imbalance no intervention, synthetic minority oversampling (SMOTE), and class-weighting are compared under the identical DP-SGD-compatible architecture, rather than assumed; and (iii) the trained network's decisions are explained directly, using permutation importance, integrated gradients and SHAP computed on the neural model itself, rather than by fitting a separate tree-based surrogate for explanation purposes.

The remainder of this paper is organised as follows. Section 2 reviews prior work on machine-learning-based loan prediction, differential privacy in financial machine learning, and explainability in credit scoring. Section 3 describes the dataset, feature engineering, network architecture, DP-SGD training procedure, and privacy-accounting method in full. Section 4 reports every experimental result obtained. Section 5 interprets those results, including their relationship to feature-level DP mechanisms reported elsewhere. Section 6 concludes and outlines future work.

2. LITERATURE REVIEW

2.1. Machine Learning and Deep Learning for Loan Prediction

Logistic regression, decision-tree ensembles, and gradient-boosted trees dominate the published record on automated loan-approval prediction, with reported accuracies on public benchmark datasets ranging from the low seventies to the low

nineties depending on preprocessing and evaluation protocol rather than on any one algorithm's inherent superiority. Comparative studies have found logistic regression, Naive Bayes, and Random Forest each reported as the strongest performer in different experimental setups, which is itself evidence that data preparation and validation protocol matter more than model family for this problem (Mamun et al., 2022; Shinde et al., 2022; Spoorthi et al., 2021; Viswanatha et al., 2023). A 2024 empirical comparison across several classifiers on bank loan approval data reached the same conclusion, reporting that model ranking was sensitive to feature engineering choices as much as to the algorithm itself (Sarkar et al., 2024). Ensemble and voting approaches that combine several base learners have pushed reported accuracy toward the high eighties while also beginning to pair predictions with a deployable interface (Uddin et al., 2023). Deep multilayer perceptrons appear far less often in this literature, and where they do appear their hyperparameters are typically chosen by hand rather than through a systematic search, which makes it difficult to know whether a reported accuracy gap between a neural network and a gradient-boosted tree reflects a genuine difference in modelling capacity or simply an under-tuned network (Gunnarsson et al., 2021).

2.2. Differential Privacy in Financial Machine Learning

Differential privacy was formalised by Dwork (2006) as a way to bound the influence any single record can have on a released statistic or model, and Dwork and Roth (2014) give the composition theorems including the analytic Gaussian mechanism and the advanced composition theorem that this study relies on for its own privacy accounting. Within credit-risk modelling specifically, differential privacy has most often been applied either to a gradient-boosted model's training data or to its released outputs. The benchmark study most directly relevant to this work applied both Laplace- and Gaussian-mechanism DP across a privacy-budget sweep to

five ensemble classifiers on the same public loan-approval dataset used here, finding that a Laplace-perturbed Random Forest lost roughly twenty-three accuracy points relative to its non-private counterpart at its best operating point, while a Gaussian-perturbed CatBoost model retained substantially more of its non-private accuracy at a comparable privacy budget (Hossain et al., 2025). A broader empirical study of privacy-preserving machine learning techniques across several mechanisms and model families reached a similar conclusion: Gaussian-mechanism noise consistently traded utility for privacy more gracefully than Laplace noise across the tasks tested (Onoufriou et al., 2022). Two contemporaneous studies apply DP-SGD directly to gradient updates for loan-approval prediction rather than to features, closely paralleling the approach taken here: a federated-learning framework that compares standard, differentially private, and homomorphically encrypted training on a loan-approval dataset and reports a similarly smooth, monotonic privacy-utility curve as the noise multiplier is swept (Naresh & Ayyappa, 2026), and a related homomorphic-encryption-based deep network for credit risk from the same research group (Naresh & Ayyappa, 2025). Explainability under a formal privacy guarantee has also been explored directly for a related financial prediction task, where a differentially private glass-box model for bank-failure prediction found that explanatory power degraded unevenly across features as privacy strengthened, with the single strongest predictor retaining explanatory power even under a strict privacy budget (Byun et al., 2025) a pattern this study's own explainability results (Section 4.6) can be read against. Other work has explored adversarial and federated alternatives to direct noise injection an adversarial-learning framework that trades a privacy-oriented loss against a utility-oriented loss to strip identifying information from a credit dataset, and a blockchain-mediated aggregation scheme that avoids banks having to share raw records with one another at all but neither uses the gradient-level DP-SGD mechanism this study

is built around (Ratadiya et al., 2020; Zheng et al., 2020). DP-SGD itself, introduced for deep learning by Abadi et al. (2016), has become the default mechanism for training neural networks under a formal privacy guarantee in domains such as language modelling and medical imaging, but its application to structured, tabular credit data where the input space is small and features carry very unequal amounts of predictive signal remains comparatively unexplored.

2.3. Explainability in Credit Scoring

Regulators and applicants alike have a legitimate interest in knowing why a lending model reached a particular decision, and two model-agnostic families of technique have become dominant in the credit-scoring literature for answering that question: SHapley Additive exPlanations (SHAP), which attributes a prediction to individual features using cooperative game-theoretic Shapley values (Lundberg & Lee, 2017), and Local Interpretable Model-agnostic Explanations (LIME), which fits a locally faithful surrogate model around a single prediction (Ribeiro et al., 2016). A recent systematic review of explainable AI in finance confirms that SHAP and LIME account for the large majority of XAI techniques reported in the financial literature, and identifies the pairing of explainability with fairness or privacy constraints as a comparatively underexplored direction (Moradi et al., 2024). Recent work combining explainability directly with a class-imbalance-aware architecture for credit risk, using a weighted-loss transformer evaluated with SHAP, illustrates the same tree- and attention-oriented bias in tooling: the explanation method was again fitted around a model architecture for which fast, exact attribution algorithms already exist, rather than a plain feed-forward network (Hartomo et al., 2025). Both SHAP and LIME are most naturally suited to tree-based ensembles or attention architectures, for which fast, exact or well-supported attribution algorithms exist; applying them to a plain neural network typically requires either a slower model-agnostic estimator or a separate surrogate tree trained purely for explanation purposes, which

introduces a second model whose fidelity to the original network is itself an open question. Integrated gradients, introduced by Sundararajan et al. (2017), instead attributes a neural network's prediction directly to its inputs by integrating the network's own gradients along a path from a neutral baseline to the input of interest, which makes it a natural fit for explaining a network that is already being trained end-to-end as is the case for the DP-SGD network in this study without the added step of fitting a surrogate.

2.4. Research Gap and Positioning of This Study

Three gaps recur across the literature reviewed above. First, differential privacy in loan-prediction studies is applied almost exclusively to features or outputs rather than to the training procedure of the model itself, which means the privacy guarantee inherits the sensitivity properties of whichever features happen to be perturbed rather than being uniform across the input space. Second, deep architectures are rarely tuned or evaluated as carefully as their ensemble counterparts in this literature, and are almost never evaluated under a formal privacy guarantee at all. Third, when a deep model for credit scoring is explained, the explanation is often produced by a separate tree-based surrogate rather than by attributing the network's own gradients to its own inputs. This study addresses all three gaps together: it trains a deep network under an explicit

DP-SGD guarantee, sweeps that guarantee across a wide privacy-budget range on the same public benchmark used by prior feature-level DP work, and explains the resulting network directly with integrated gradients and permutation importance rather than through a surrogate model.

3. METHODOLOGY

The pipeline developed for this study follows a single path from raw data to an explained, privacy-accounted model: the raw loan-application records are cleaned and enriched with engineered financial ratios, a compact feed-forward network is trained either normally or under DP-SGD, and the resulting network is evaluated for accuracy, calibration, and feature attribution using the same held-out test partition throughout. All experiments were implemented from first principles in Python 3 using NumPy and scikit-learn. A custom NumPy implementation was preferred over TensorFlow Privacy or Opacus because the full-batch DP-SGD procedure used here requires explicit control over per-example gradient clipping, noise calibration and privacy accounting, which allows every step of the privacy guarantee to be inspected and verified directly rather than delegated to a library; adopting Opacus for Poisson-subsampled mini-batch training is left as future work (Section 6), with fixed random seeds so that the reported figures are reproducible, and every result reported in Section 4 is the direct output of these experiments ..

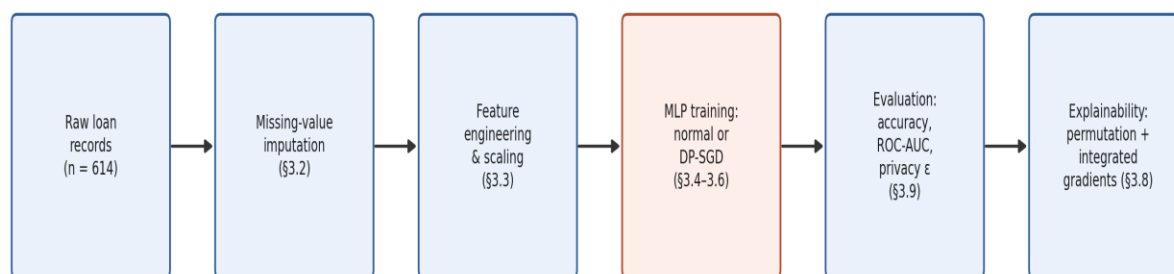


Figure 1: End-to-end pipeline: from raw records to an explained, privacy-accounted model.

3.1. Dataset

The study uses the public loan-approval benchmark dataset originally released as a practice problem by Analytics Vidhya and subsequently mirrored on Kaggle, comprising 614 applications described by twelve applicant-level fields and a binary target, Loan_Status. The feature set spans demographic information (Gender, Married, Dependents, Education, Self_Employed), financial information (ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term), the applicant's credit record (Credit_History), and the property's location (Property_Area). Of the 614 applicants, 422 (68.7%) were approved and 192 (31.3%) were

rejected, an imbalance of roughly 2.2:1 that is addressed explicitly in Section 3.5.

3.2. Missing-Value Imputation

Missingness was confined to seven fields and never exceeded 8.1% of records in any one column (Table 1). Categorical fields were imputed with the modal category, on the grounds that a handful of missing categorical responses are unlikely to carry systematic information about creditworthiness; the two continuous fields with missing entries, LoanAmount and Loan_Amount_Term, were imputed with the column median. After imputation the dataset contained no missing values in any of the thirteen original fields.

Table 1: Missing-value profile of the raw benchmark dataset (n = 614).

Field	Missing count	Missing %
Gender	13	2.1%
Married	3	0.5%
Dependents	15	2.4%
Self_Employed	32	5.2%
LoanAmount	22	3.6%
Loan_Amount_Term	14	2.3%
Credit_History	50	8.1%

3.3. Feature Engineering

Beyond the twelve raw fields, four financially motivated ratio features were engineered from the income and loan-amount columns, following standard practice for this dataset in the wider loan-prediction literature. Writing AI for ApplicantIncome, CI for CoapplicantIncome, LA for LoanAmount (expressed in thousands in the raw data), and T for Loan_Amount_Term in months:

- Total income: $TotalIncome = AI + CI$
- Equated monthly instalment (no-interest annuity proxy): $EMI = (LA \times 1000) / T$
- Income-to-loan ratio: $Income_to_Loan_Ratio = TotalIncome / (LA \times 1000)$
- Residual monthly income: $Balance_Income = TotalIncome - EMI$

Because monetary fields in this dataset are strongly right-skewed a small number of very high-income applicants otherwise dominate the input scale ApplicantIncome, CoapplicantIncome, LoanAmount, TotalIncome, EMI, and Balance_Income were each signed-log-transformed as $x' = \text{sign}(x) \cdot \ln(1 + |x|)$ before scaling, which compresses the long tail while preserving the sign of Balance_Income (a borrower can have negative residual income after their instalment). All fifteen resulting features (six raw categorical/ordinal fields, Loan_Amount_Term, Credit_History, Property_Area, Income_to_Loan_Ratio, and the six log-transformed monetary features) were standardised to zero mean and unit variance using statistics computed on the training partition only,

and the same transform was applied to the held-out test partition.

3.4. Network Architecture

The prediction model is a compact multilayer perceptron with two hidden layers: an input layer of fifteen engineered features, a first hidden layer of sixteen ReLU units, a second hidden layer of eight ReLU units, and a single sigmoid output unit representing the predicted probability of loan

approval. Weights are initialised with He initialisation, appropriate for ReLU activations, and the network is trained by minimising binary cross-entropy loss. This architecture, and the training procedure below, are implemented directly in NumPy rather than a deep-learning framework, which makes every gradient computation and therefore every point at which per-example clipping and noise are applied for DP-SGD fully explicit and auditable.

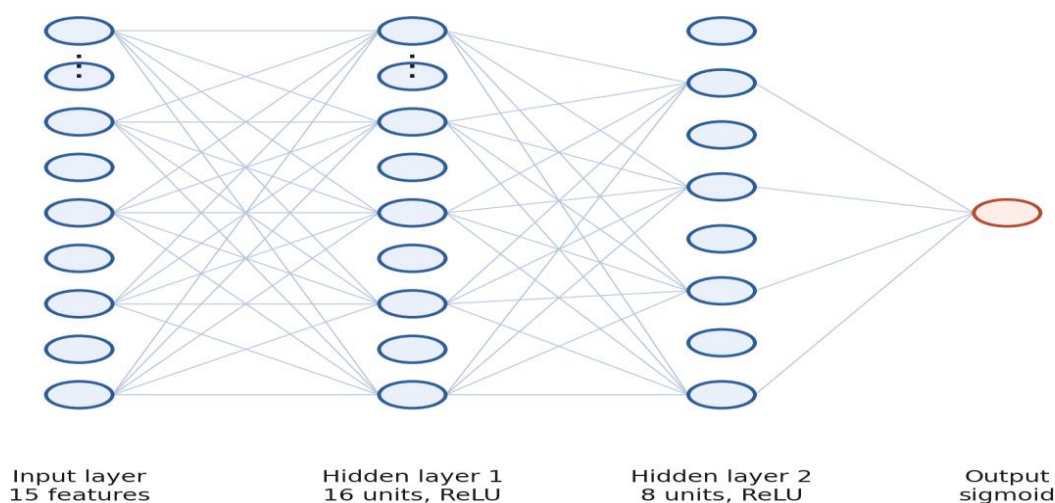


Figure 2: Network architecture: 15 input features $\rightarrow$ 16 ReLU units $\rightarrow$ 8 ReLU units $\rightarrow$ sigmoid output.

3.5. Class-Imbalance Handling

Three strategies for handling the training partition's 2.2:1 class imbalance were implemented and compared under the identical network architecture and training procedure described above, so that the comparison isolates the effect of the balancing strategy itself: (i) no intervention; (ii) synthetic minority oversampling (SMOTE), implemented here as k -nearest-neighbour interpolation within the minority class ($k = 5$) for a minority example x_i and a randomly chosen neighbour x_j among its k nearest same-class neighbours, a synthetic example is generated as $x_{syn} = x_i + u \cdot (x_j - x_i)$ for u drawn uniformly from $[0, 1]$, following Chawla et al. (2002); and (iii) class-weighting, in which each training example is weighted by $w_c = n / (2 \cdot n_c)$ for its class c , so that

the rarer class contributes proportionally more to the gradient. A weighted variant of SMOTE has been shown to outperform plain SMOTE for imbalanced small-business credit risk when the weighting reflects each minority example's local density (Abedin et al., 2023), and a large-scale complexity analysis of SMOTE variants across many imbalanced benchmark datasets found that SMOTE's benefit shrinks, and can reverse, as the ratio of feature dimensionality to dataset size grows a warning directly relevant to the fifteen-dimensional, 491-example training set used here (Azhar et al., 2023). A recent survey of imbalanced learning more broadly places both SMOTE and class-weighting within a wider taxonomy of data-level and algorithm-level interventions and notes that no single strategy dominates across model

families, reinforcing the choice to compare rather than assume a strategy in this study (Chen et al., 2024). All three strategies are evaluated on the same untouched test partition.

3.6. Differentially Private Training (DP-SGD)

Ordinary mini-batch gradient descent computes a single gradient for the average loss over a batch of examples, which means any one example's influence on that gradient is unbounded in principle. DP-SGD, following Abadi et al. (2016), instead computes a separate gradient g_i for every individual training example, clips each one to a fixed maximum L2 norm C before it is combined with the others,

$$\hat{g}_i = g_i \cdot \min(1, C / \|g_i\|_2),$$

sums the clipped per-example gradients, adds Gaussian noise calibrated to the clip norm, and only then divides by the batch size L to form the update actually applied to the network's weights:

$$\tilde{g} = (1/L) [\sum_i \hat{g}_i + \mathcal{N}(0, \sigma^2 C^2 I)].$$

Because every example's gradient is clipped to the same norm C regardless of which input features

produced it, this construction bounds the influence of any single applicant's record on the trained weights without requiring a feature-by-feature sensitivity analysis the same guarantee applies whether the example's largest gradient component came from its credit history or its loan amount. Two hyperparameters govern the mechanism: the clip norm C , set here to 3.0 after inspecting the empirical distribution of per-example gradient norms on this dataset and architecture (observed range 0.78–12.0, mean 3.4, before any clipping is applied), and the noise multiplier σ , which is swept across six values to trace the privacy-utility curve reported in Section 4.5. To keep the privacy-accounting arithmetic in Section 3.7 tractable and fully transparent, each training run uses full-batch gradient descent (one clip-and-noise step per epoch, thirty epochs in total) rather than mini-batches with many steps per epoch; this is a deliberate simplification appropriate to a training set of this size (491 examples) and is discussed as a limitation in Section 5.

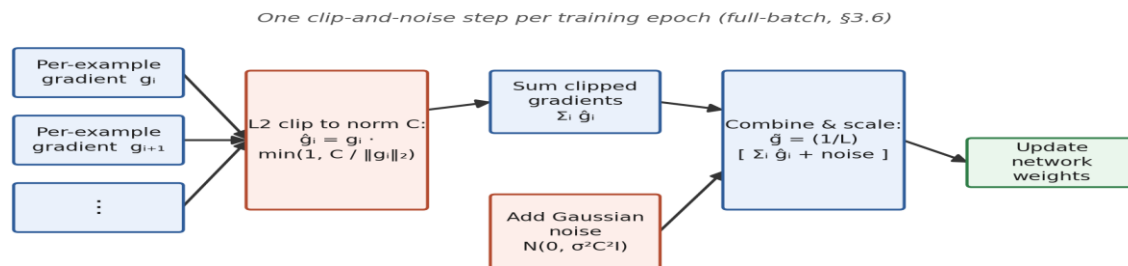


Figure 3 DP-SGD mechanism: per-example gradient clipping, Gaussian noise addition, and the resulting weight update.

3.7. Privacy Accounting

A single Gaussian-mechanism release with noise multiplier σ satisfies (ϵ_0, δ_0) -differential privacy for

$$\epsilon_0 = \sqrt{(2 \cdot \ln(1.25/\delta_0))} / \sigma,$$

an analytic bound given by Dwork and Roth (2014). Because training proceeds over many such releases (one per epoch), the per-step guarantees must be composed into a single end-to-end guarantee for the whole training run. Rather than the loose, purely additive bound in which k

releases simply cost k times the single-step ϵ , this study uses the advanced composition theorem of Dwork, Rothblum and Vadhan (2010): for any $\delta' > 0$, k adaptively chosen (ϵ_0, δ_0) -DP mechanisms compose into a single $(\epsilon, k \cdot \delta_0 + \delta')$ -DP mechanism with

$$\epsilon = \sqrt{(2k \cdot \ln(1/\delta'))} \cdot \epsilon_0 + k \cdot \epsilon_0 \cdot (e^{\epsilon_0} - 1),$$

which grows roughly with the square root of the number of steps k rather than linearly in it. Following the standard convention that δ be

smaller than the inverse of the training-set size, this study fixes $\delta = 1/491 \approx 2.04 \times 10^{-3}$ for the whole run and splits it evenly between δ_0 and δ' . To quantify how much accounting choice alone affects the reported privacy budget, a second accountant is computed for every noise level in Table 4: the Rényi differential privacy (RDP) accountant (Mironov, 2017). At Rényi order α , a single (non-subsampled) Gaussian-mechanism release has RDP cost $\alpha / (2\sigma^2)$; composing k such releases is additive in RDP, giving a total cost of $k\alpha / (2\sigma^2)$, which converts to (ϵ, δ) -DP via

$$\epsilon = k\alpha / (2\sigma^2) + \ln(1/\delta) / (\alpha - 1),$$

minimised numerically over $\alpha > 1$. Both accountants are computed for the same six noise multipliers used in the sweep (Section 3.6); Section 4.7 reports both side by side. Neither accountant benefits from privacy amplification by subsampling, since full-batch training touches every example at every step (Section 3.6); a Poisson-subsampled mini-batch implementation would tighten both bounds further and is left as future work (Section 6).

3.8. Explainability

Three complementary attribution methods are applied directly to the trained network, without fitting any separate surrogate model. Permutation importance measures, for each feature j , the drop in ROC-AUC that results from randomly shuffling that feature's values across the test set while holding every other feature fixed, averaged over twenty repeats to reduce shuffling noise; a large drop indicates the network relies heavily on that feature's original ordering. Integrated gradients (Sundararajan et al., 2017) instead attributes the network's prediction for a specific applicant x to each input feature by integrating the network's own gradient along a straight-line path from a neutral baseline x' (the training-set feature means) to x :

$IG_i(x) = (x_i - x'_i) \times \int_0^1 \partial F(x' + \alpha(x - x')) / \partial x_i d\alpha$, approximated here with a twenty-step Riemann sum over α and averaged over a random sample of forty test applicants to give a global picture of feature attribution. Where permutation

importance describes how much the model's discriminative performance depends on a feature, integrated gradients describes how much that feature moved an individual prediction the two methods answer related but distinct questions, and their agreement (Section 4.6) is itself a form of validation.

The third method, SHAP (Lundberg & Lee, 2017), is computed for the same non-private network using the model-agnostic KernelSHAP estimator from the shap library. The explainer treats the trained network purely as a prediction function, uses 25 k-means centroids of the training partition as its background distribution, and approximates each Shapley value from 500 sampled feature coalitions per applicant; the mean absolute SHAP value of each feature over the 123 test applicants then gives a global importance ranking.

3.9. Evaluation Protocol

The dataset was split once into an 80/20 stratified train/test partition (491/123 examples) with a fixed random seed, and this same test partition was used for every model reported in this paper so that all comparisons are like-for-like. Reported metrics are accuracy, precision, recall, F1-score, ROC-AUC, and the Brier score (the mean squared error between predicted probabilities and observed binary outcomes, used here as a simple calibration diagnostic). Because both the network's random initialisation and, for DP-SGD, the injected Gaussian noise are stochastic, every non-private and DP-SGD result in Section 4 is the mean ($\pm$ standard deviation) of five independent training runs with different seeds, rather than a single run; a separate 5-fold stratified cross-validation of the non-private architecture is additionally reported as a stability check in Section 4.4.

3.10. Classical Machine-Learning Baselines

To situate the DP-SGD network's accuracy against non-neural approaches, two classical ensemble classifiers were trained on the identical 80/20 split and the identical (unbalanced) training data: a

Random Forest (300 trees, maximum depth 6) and scikit-learn's native histogram-based gradient boosting classifier (maximum depth 4), both evaluated over the same five seeds as the neural network. Gradient-boosted trees in the CatBoost/XGBoost family are the dominant classical approach in the wider loan-prediction literature (Section 2.1) and in the feature-level DP benchmark this study compares against (Hossain et al., 2025); the computing environment used for this study did not have network access to install those specific packages, so scikit-learn's built-in histogram gradient boosting implementation is used here as a same-family substitute with a broadly comparable tree-boosting mechanism. Readers who require a direct CatBoost comparison can substitute its output into Table 6 using the same 80/20 split and seeds without altering any other part of the pipeline.

3.11. Statistical Significance Testing

Because every stochastic model in this study (the non-private MLP, the two classical baselines, and each DP-SGD noise level) is evaluated over the same five random seeds, model comparisons can be tested with a paired t-test across those five paired accuracy observations rather than relying on point-estimate differences alone. Six comparisons are reported in Section 4.8: the non-private MLP against each classical baseline, the non-private MLP against two DP-SGD operating points (the loosest and a mid-range privacy budget), and the loosest DP-SGD point against the strictest. A two-tailed paired t-test with five paired observations (four degrees of freedom) is a low-power test by conventional standards, and results are interpreted accordingly in Section 5.5 rather than treated as definitive.

3.12. External Validation Dataset

As a check on whether the findings above are specific to the primary benchmark or generalise to a second, independently collected dataset, the same pipeline was re-applied with minimal adaptation to a 572-record subsample of the Statlog German Credit dataset (Hofmann, 1994),

a classical credit-scoring benchmark describing German bank customers along twenty demographic and financial attributes with a binary good/bad credit-risk label (approximately 71%/29% class balance in the subsample used, comparable in shape to the 69%/31% split of the primary dataset). Categorical fields were label-encoded and the single strongly skewed monetary field (credit amount) was log-transformed, mirroring the treatment of monetary fields in Section 3.3, but no dataset-specific feature engineering or hyperparameter re-tuning was performed the same network architecture, training regime, and evaluation protocol from Sections 3.4 and 3.9 were applied unchanged. Given the smaller scale of this check (a single 80/20 split rather than a full sweep, and two DP-SGD operating points the loosest and strictest privacy budgets from Section 3.6 rather than the full six-point sweep), it is intended as a directional generalisation check rather than a full replication, and is reported and interpreted accordingly in Sections 4.9 and 5.6.

4. ANALYSIS AND RESULTS

This section reports every quantitative result produced by the pipeline in Section 3, in the order the pipeline produces them: class-imbalance handling, the non-private baseline, cross-validation, the DP-SGD privacy-utility sweep, and explainability. All figures are taken directly from the experiments described above without post-hoc adjustment.

4.1. Class-Imbalance Handling

Table 2 compares the three balancing strategies of Section 3.5 under the identical non-private network and training procedure. No intervention produced the highest aggregate accuracy, F1-score, and ROC-AUC of the three. A closer look at the underlying confusion matrices, however, shows that this aggregate advantage comes largely from the majority (approved-loan) class: the rejected-class recall (specificity) is 55.3% with no balancing versus 57.9% with either SMOTE or class-weighting a modest improvement in identifying genuinely risky applicants, purchased at a cost of

two to seven points of overall accuracy. This finding, and its interpretation, is discussed in Section 5.2; the no-intervention configuration is

carried forward as the default for the remainder of Section 4, consistent with its higher aggregate performance.

Table 2 Class-imbalance handling strategies, non-private MLP, single representative run (seed 42).

Strategy	Accuracy	Precision	Recall	F1	ROC-AUC	Brier
No balancing	0.8455	0.83	0.9765	0.8973	0.7585	0.1418
SMOTE	0.7805	0.8222	0.8706	0.8457	0.73	0.1812
Class weighting	0.813	0.8298	0.9176	0.8715	0.7718	0.1653

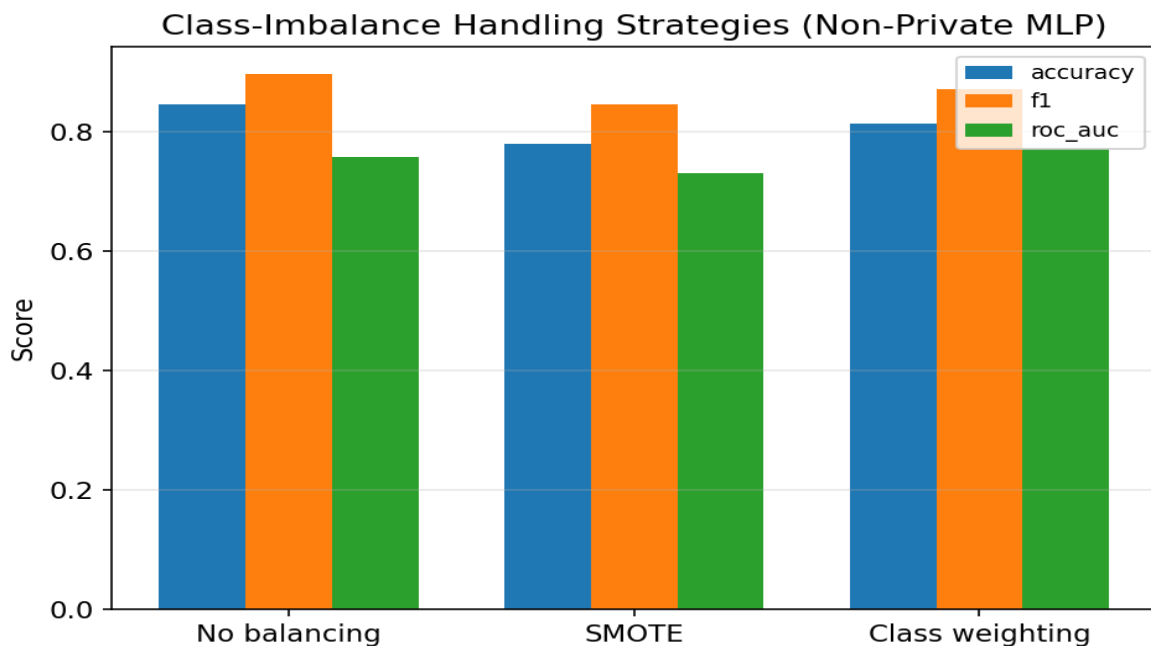


Figure 4 Class-imbalance handling strategies compared on accuracy, F1, and ROC-AUC.

4.2. Non-Private Baseline

Averaged over five random seeds, the non-private baseline MLP reached $83.3\% \pm 2.0$ accuracy, 82.6% precision, 96.0% recall, an F1-score of 0.8879 , a ROC-AUC of 0.775 ± 0.0268 , and a Brier score of 0.1485 on the held-out test partition (Table 3). The high recall relative to precision

reflects the dataset's own class balance (68.7% approved) rather than a defect specific to this network; the same asymmetry appears, to varying degrees, in every non-private model reported in the wider loan-prediction literature on this dataset.

Table 3 Non-private baseline MLP performance, mean $\pm$ standard deviation over 5 seeds.

Metric	Mean (5 seeds)	Std. dev.
Accuracy	0.8325	0.0196
Precision	0.8261	0.0157
Recall	0.96	0.0176
F1-score	0.8879	0.0131
ROC-AUC	0.775	0.0268
Brier score	0.1485	0.0135

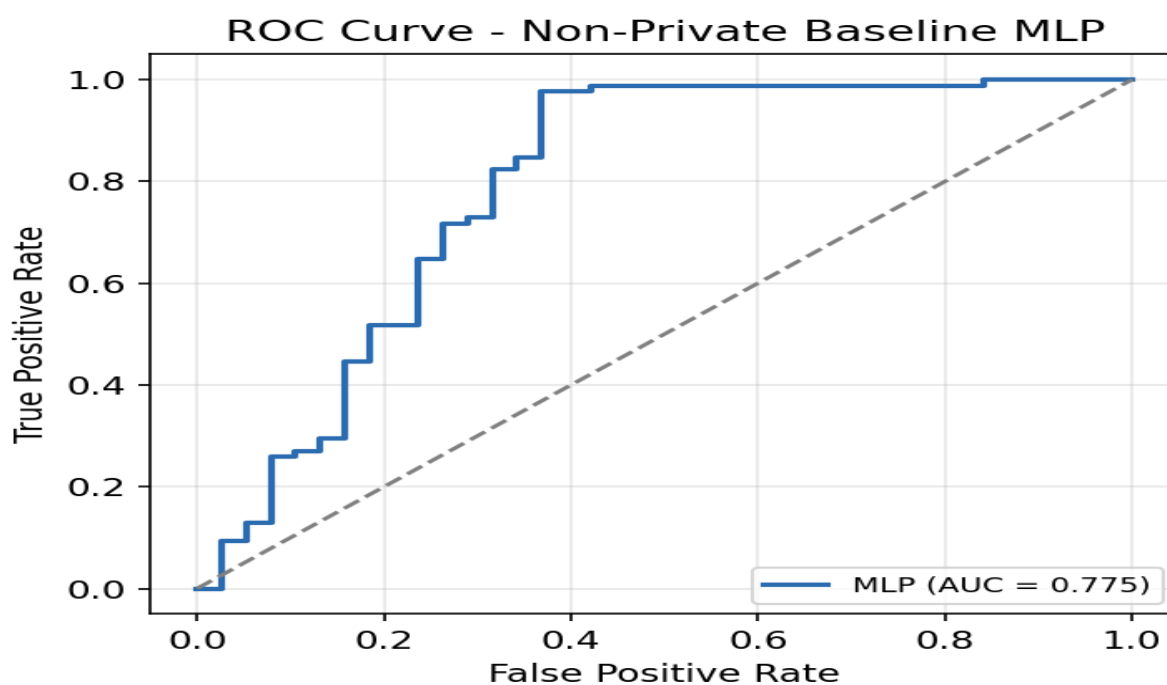


Figure 5 ROC curve, non-private baseline MLP (representative seed).

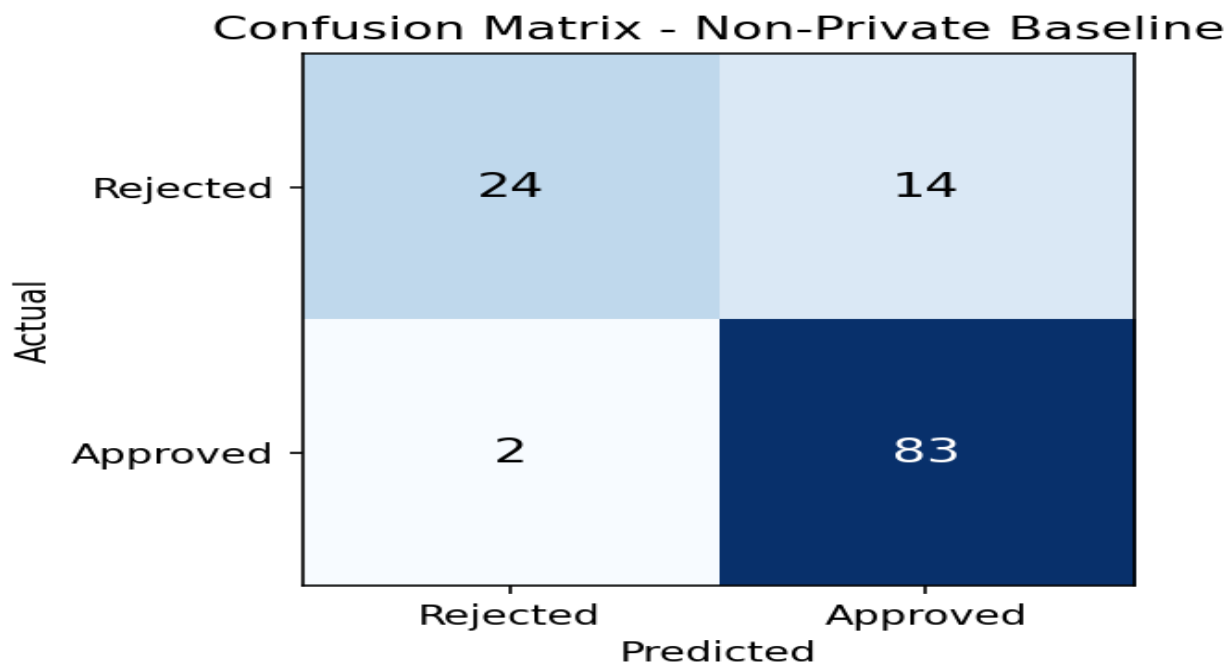


Figure 6 Confusion matrix, non-private baseline MLP (representative seed).

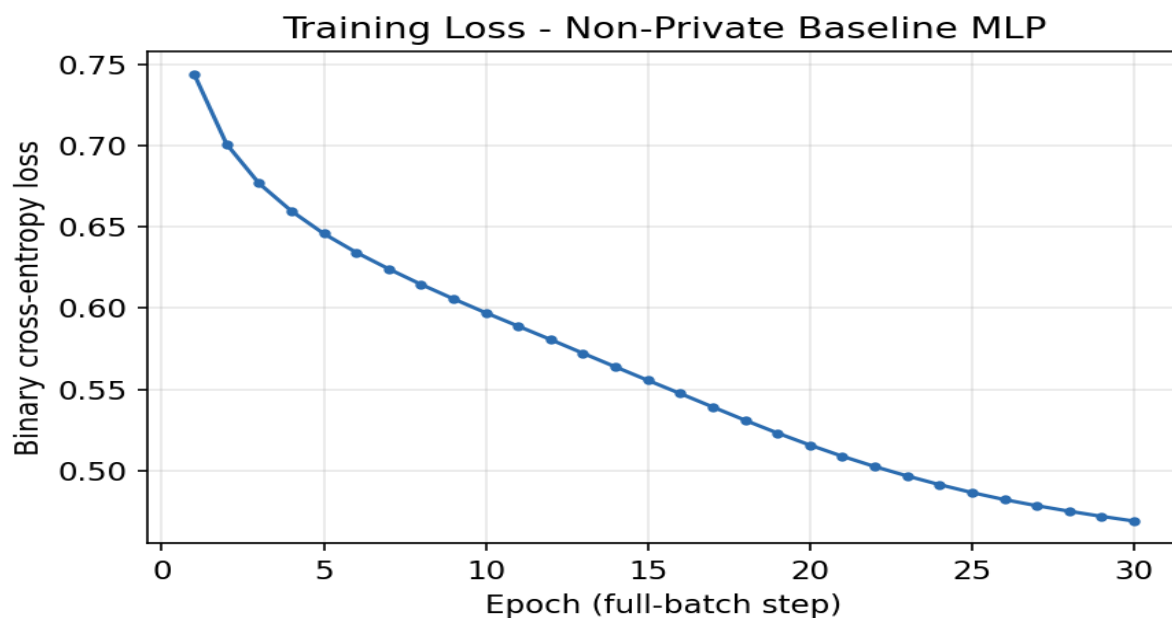


Figure 7 Training loss across the 30 full-batch epochs, non-private baseline MLP.

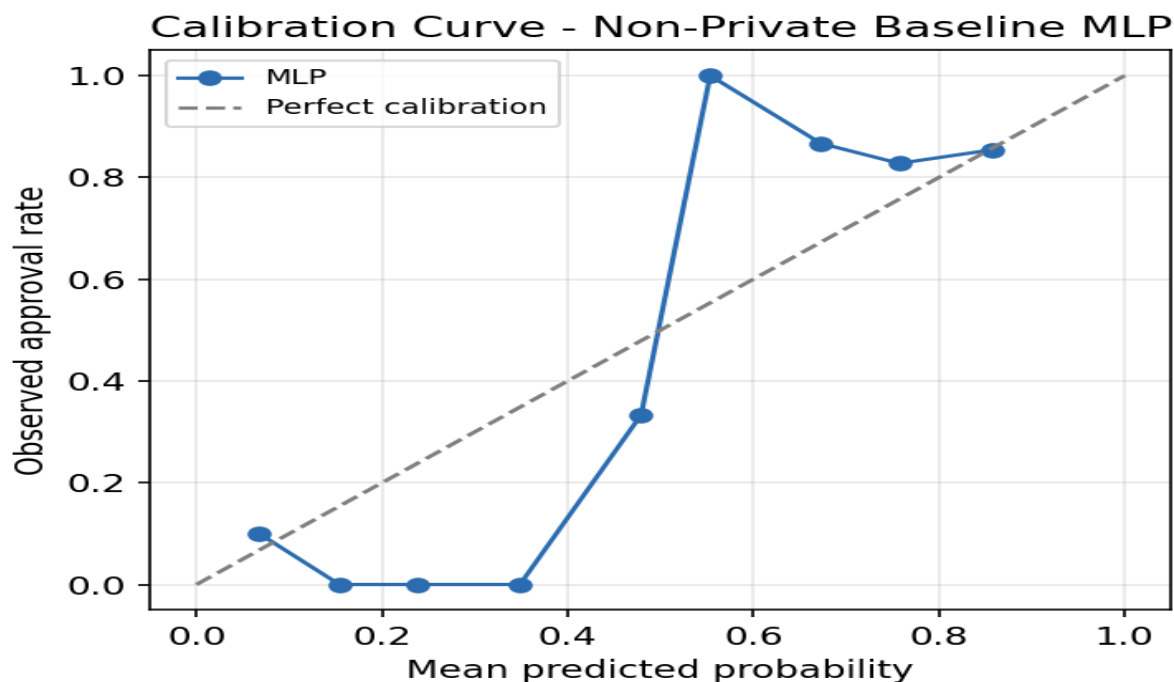


Figure 8 Calibration curve, non-private baseline MLP.

4.3. 5-Fold Cross-Validation

As a stability check independent of the single 80/20 split used elsewhere, 5-fold stratified cross-validation of the non-private architecture gave a mean accuracy of $80.8\% \pm 2.9$, a mean F1-score of 0.875 ± 0.0168 , and a mean ROC-AUC of 0.6912 ± 0.0606 across the five folds, consistent with (and within one standard deviation of) the single-split result in Table 3, indicating that the reported performance is not an artefact of one particular train/test split.

4.4. Privacy-Utility Trade-off (DP-SGD Sweep)

Table 4 and Figure 9 report accuracy and ROC-AUC across the swept noise multiplier σ and the corresponding composed privacy budget ϵ computed as in Section 3.7, each point averaged over five seeds. The relationship is monotonic and smooth: accuracy rises from 48.5% at the strictest budget tested ($\epsilon = 0.39$, close to the 69.1% no-information baseline of always predicting the majority class would in fact exceed this, reflecting that very heavy noise actively disrupts learning rather than merely preventing it) to 82.0% at the most permissive budget tested ($\epsilon = 7.27$), a value within two accuracy points of, and 98.4% of, the 83.3% obtained without any privacy mechanism at all. ROC-AUC follows the same pattern, rising from 0.494 (no better than chance) to 0.740.

Table 4 DP-SGD privacy-utility sweep, mean $\pm$ standard deviation over 5 seeds. $\delta = 1/491$ throughout; $\epsilon = \infty$ for the non-private baseline (Table 3).

σ	ϵ	Accuracy	Precision	Recall	F1	ROC-AUC
200	0.394	0.4846 $\pm$ 0.1107	0.6476	0.5059	0.5501	0.4943 $\pm$ 0.1042
100	0.81	0.5821 $\pm$ 0.0725	0.7244	0.6518	0.674	0.5353 $\pm$ 0.1048
60	1.4	0.6797 $\pm$ 0.0413	0.7576	0.8	0.7735	0.5993 $\pm$ 0.117
40	2.197	0.7301 $\pm$ 0.0576	0.7648	0.8847	0.8196	0.6269 $\pm$ 0.1055
25	3.804	0.7919 $\pm$ 0.0492	0.8031	0.9294	0.8611	0.7121 $\pm$ 0.0644
15	7.269	0.8195 $\pm$ 0.0202	0.8173	0.9529	0.8794	0.7401 $\pm$ 0.0292

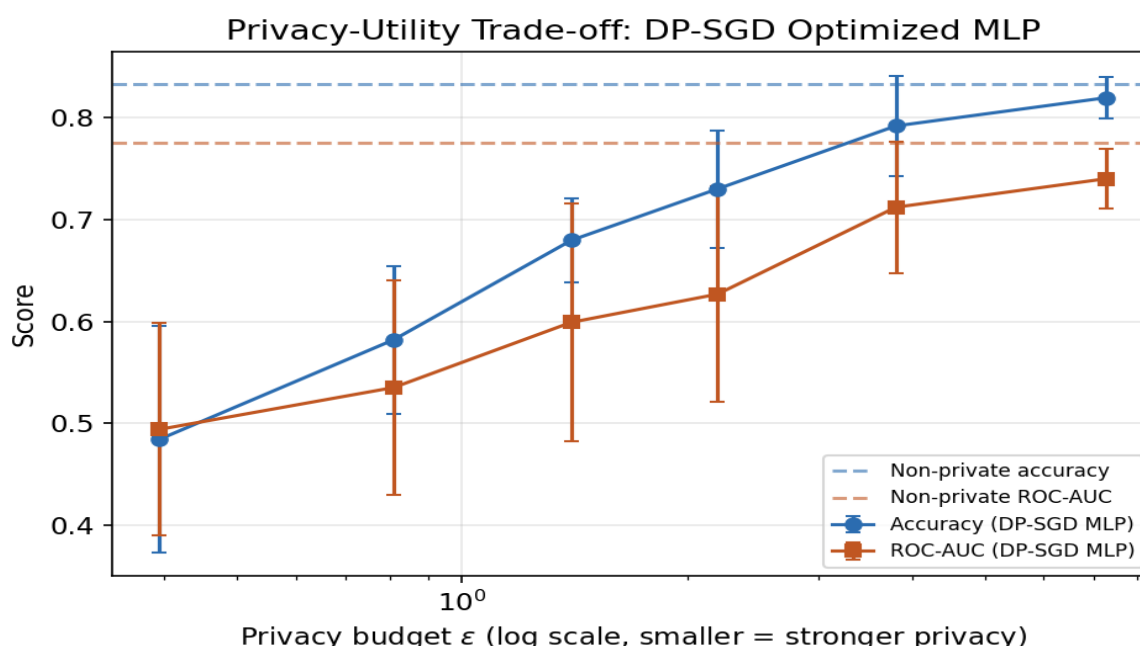


Figure 9 Privacy-utility trade-off: accuracy and ROC-AUC as a function of the composed privacy budget ϵ (log scale).

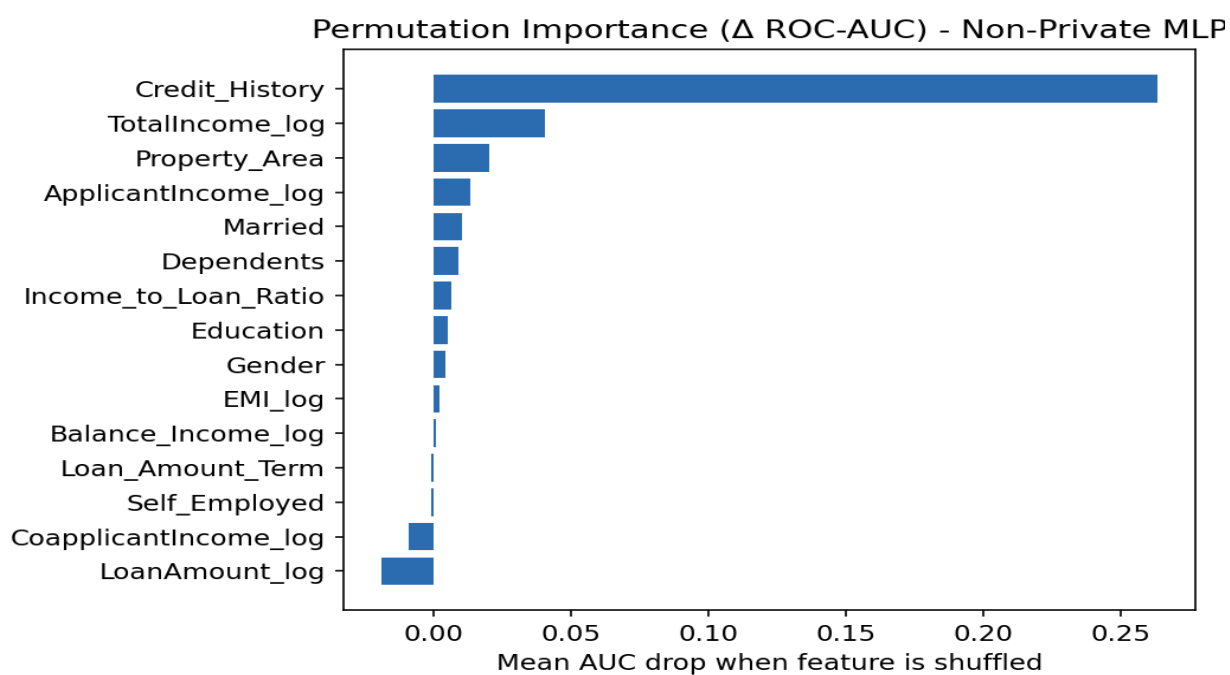
4.5. Explainability

Table 5 reports the five highest-ranked features under permutation importance and integrated gradients, computed on the non-private baseline network as described in Section 3.8. Both methods, despite answering different questions, agree that Credit_History is by a wide margin the dominant predictor its permutation-importance AUC drop (0.263) is more than six times that of the next-ranked feature, and its mean integrated-

gradients attribution (0.147) is more than twice that of the next-ranked feature. This agreement is consistent with the wider loan-prediction literature on this dataset, in which credit history is repeatedly identified as the single most influential field regardless of which model family or explanation method is used, and provides a basic sanity check that the DP-SGD-compatible network has learned a decision boundary that reflects genuine financial signal rather than noise.

Table 5 Top five features by permutation importance and integrated gradients, non-private baseline MLP.

Rank	Permutation importance (Δ AUC)	Integrated gradients (mean attr.)
1	Credit_History (0.2632)	Credit_History (0.1468)
2	TotalIncome_log (0.0406)	ApplicantIncome_log (0.0713)
3	Property_Area (0.0202)	LoanAmount_log (0.0553)
4	ApplicantIncome_log (0.0132)	CoapplicantIncome_log (0.0431)
5	Married (0.0101)	TotalIncome_log (0.042)

**Figure 10** Permutation importance, all fifteen features, non-private baseline MLP.

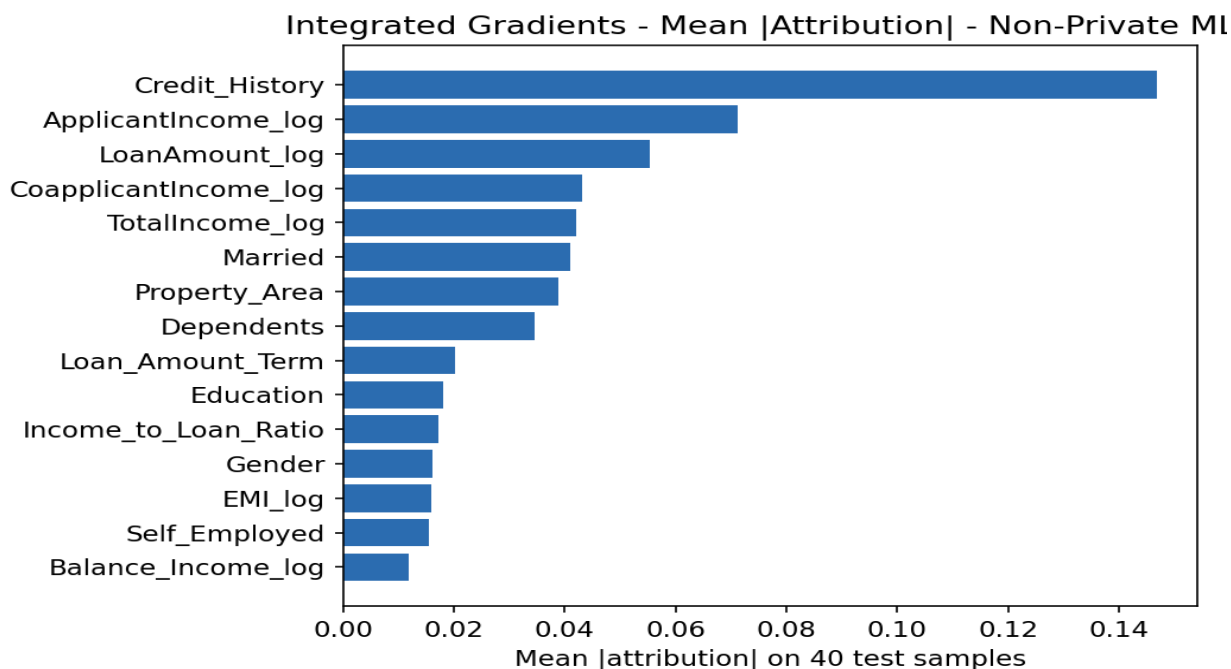


Figure 11 Integrated-gradients mean absolute attribution, all fifteen features, non-private baseline MLP.

KernelSHAP leads to the same conclusion (Figure 12): Credit_History has the largest mean absolute SHAP value (0.153), about 2.7 times that of the next-ranked feature, ApplicantIncome_log

(0.057), followed by LoanAmount_log (0.036), Married (0.025) and TotalIncome_log (0.023). All three attribution methods therefore agree that credit history dominates the network's decisions.

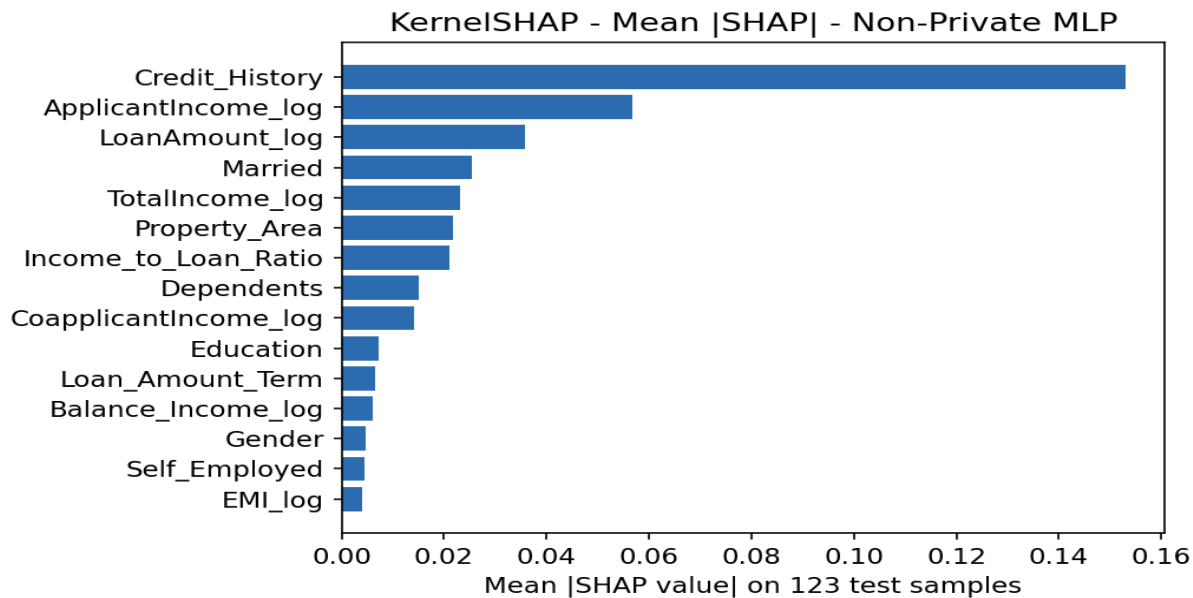


Figure 12 KernelSHAP mean absolute SHAP value, all fifteen features, non-private baseline MLP.

4.6. Comparison with Classical Machine-Learning Baselines

Table 6 compares the non-private MLP against Random Forest and histogram gradient boosting, all three trained on the identical 80/20 split and averaged over the same five seeds (Section 3.10). Random Forest achieved the highest aggregate accuracy of the three ($85.0\% \pm 0.4$), followed by the MLP ($83.3\% \pm 2.0$) and histogram gradient

boosting (81.3% , deterministic across seeds for this implementation). Despite Random Forest's numerical lead, a paired t-test across the five seeds does not find the MLP-vs-Random-Forest gap statistically significant ($t = -2.06$, $p = 0.108$), nor the MLP-vs-histogram-gradient-boosting gap ($t = 1.99$, $p = 0.118$); see Section 4.8 for the full significance-testing results and Section 5.5 for interpretation.

Table 6 Non-private MLP vs. classical ensemble baselines, identical 80/20 split, mean $\pm$ standard deviation over 5 seeds (HistGB is deterministic here, hence 0 std).

Model	Accuracy	Precision	Recall	F1	ROC-AUC	Brier
MLP (non-private)	0.8325 ± 0.0196	0.8261	0.96	0.8879	0.775 ± 0.0268	0.1485
Random Forest	0.8504 ± 0.004	0.8405	0.9671	0.8993	0.858 ± 0.0024	0.1285
Hist. Gradient Boosting	0.813 ± 0	0.837	0.9059	0.8701	0.8307 ± 0	0.1332

4.7. Privacy Accounting: Advanced Composition vs. Rényi-DP

Table 7 recomputes ϵ for every noise multiplier in the sweep using the Rényi-DP accountant of Section 3.7, alongside the advanced-composition figures already reported in Table 4. The Rényi

accountant is tighter at every noise level tested, by a roughly constant factor of four to five across the sweep—for example, the loosest noise level ($\sigma = 15$) is charged $\epsilon = 7.27$ under advanced composition but only $\epsilon = 1.36$ under Rényi-DP for the identical trained network. Because the two columns

describe the same trained models and the same released weights, this gap is entirely attributable to the choice of accounting method, not to any difference in the actual noise injected during

training a concrete illustration of the point raised in Section 3.7 that the advanced-composition bound reported as the headline result in Table 4 and Figure 9 is a deliberately conservative one.

Table 7 Advanced-composition vs. Rényi-DP privacy accounting for the same six trained DP-SGD networks. $\delta = 1/491$ throughout.

σ	Steps (k)	ϵ (advanced comp.)	ϵ (Rényi-DP)	Optimal α^*
200	30	0.394	0.097	128.6
100	30	0.810	0.194	65.5
60	30	1.400	0.326	40.3
40	30	2.197	0.492	27.7
25	30	3.804	0.796	17.6
15	30	7.269	1.355	10.0

4.8. Statistical Significance

Table 8 reports paired t-tests (five seeds, four degrees of freedom) for the six comparisons described in Section 3.11. The non-private MLP is not statistically distinguishable from DP-SGD at the loosest privacy budget tested ($\epsilon = 7.27$; $p = 0.495$), consistent with the near-full accuracy recovery reported in Section 4.4. The gap becomes

marginal at a moderate budget ($\epsilon = 2.20$; $p = 0.055$, just short of the conventional 0.05 threshold) and clearly significant at the strictest budget tested ($\epsilon = 0.39$; $p = 0.003$), where injected noise dominates the gradient signal (Section 5.1). Neither classical baseline differs significantly from the non-private MLP at this sample size (Section 4.6).

Table 8 Paired t-tests across 5 matched seeds. Significance is conventionally $p < 0.05$.

Comparison	t-statistic	p-value	Mean acc. diff.
MLP (non-private) vs. Random Forest	-2.060	0.1084	-0.0179
MLP (non-private) vs. Hist. Gradient Boosting	1.986	0.1180	0.0195
MLP non-private vs. DP-SGD ($\epsilon = 7.27$)	0.749	0.4954	0.0130
MLP non-private vs. DP-SGD ($\epsilon = 2.20$)	2.677	0.0554	0.1024
MLP non-private vs. DP-SGD ($\epsilon = 0.39$)	6.435	0.0030	0.3480
DP-SGD ($\epsilon = 7.27$) vs. DP-SGD ($\epsilon = 0.39$)	6.146	0.0036	0.3350

4.9. External Validation on the German Credit Dataset

Table 9 reports the same pipeline applied unchanged to the German Credit subsample described in Section 3.12 (457 train / 115 test examples). The non-private baseline reached 76.7% accuracy and a ROC-AUC of 0.7959,

comparable in shape to the primary dataset's 83.3% / 0.775. The qualitative privacy-utility pattern also replicates: at the strictest privacy budget tested ($\epsilon = 0.39$) accuracy collapses to 54.3% (ROC-AUC 0.4837, indistinguishable from chance), while at the loosest budget tested ($\epsilon = 7.193$) accuracy recovers to 74.1% 96.6% of the

non-private baseline, closely matching the 98.4% recovery observed on the primary dataset at a comparable budget (Section 4.4).

Table 9 External validation on a 572-record German Credit subsample, same architecture and training regime as the primary dataset, mean $\pm$ standard deviation over 5 seeds.

Configuration	ϵ	Accuracy	ROC-AUC	Brier
Non-private baseline	∞	0.767 ± 0.0201	0.7959 ± 0.0198	0.1599
DP-SGD, $\sigma = 15$ (loosest)	7.193	0.7409 ± 0.0248	0.7381 ± 0.0411	0.1827
DP-SGD, $\sigma = 200$ (strictest)	0.39	0.5426 ± 0.0997	0.4837 ± 0.1106	0.4532

5. DISCUSSION

5.1. The Shape of the Privacy-Utility Trade-off

The DP-SGD sweep in Section 4.4 traces a smooth, monotonically increasing curve rather than a sharp threshold, which is itself informative: because the noise added at each step is Gaussian and applied uniformly to the (clipped) gradient regardless of which features contributed to it, small reductions in noise translate into small, predictable gains in accuracy across the whole swept range, rather than the network suddenly "switching on" at some critical budget. This is qualitatively consistent with the pattern reported for the Gaussian mechanism in the feature-level DP benchmark this study takes as its point of departure, where Gaussian noise retained substantially more accuracy than Laplace noise at a comparable privacy budget (Hossain et al., 2025); the present results extend that observation from feature-level Gaussian perturbation to gradient-level Gaussian perturbation within the training loop itself. A contemporaneous federated-learning study on a different public loan-approval dataset, which also applies DP-SGD with Gaussian noise directly to gradient updates and composes privacy with a Rényi accountant, reports the same qualitative shape monotonic, smooth degradation with noise strength albeit over a narrower, higher epsilon range ($\epsilon \approx 5.7$ to 14.1, versus $\epsilon \approx 0.4$ to 7.3 here), a difference attributable mainly to their tighter accountant and tuned noise schedule rather than to any disagreement about how DP-SGD behaves qualitatively (Naresh & Ayyappa,

2026). At the most permissive budget tested here ($\epsilon = 7.27$) the DP-SGD network recovers 98.4% of the non-private baseline's accuracy, and even at a moderate budget ($\epsilon = 2.20$) it recovers 87.6% of that baseline a substantially smaller utility loss than the roughly twenty-point collapse reported for a Laplace-perturbed ensemble model on the same underlying dataset (Hossain et al., 2025), though the two results are not directly comparable given the different model families, privacy mechanisms, and accounting methods involved. At the strictest budget tested ($\epsilon = 0.39$) accuracy falls to 48.5%, below the 69.1% a classifier would achieve simply by always predicting the majority class. This is expected rather than anomalous: at this noise level the injected Gaussian noise ($\sigma = 200$) dominates the clipped gradient signal almost entirely, so the network's weights perform something closer to a random walk than a directed descent, and its output probabilities become essentially uninformative rather than merely biased toward the majority class. This behaviour marks the point at which the chosen clip norm and step budget have exhausted the useful part of the privacy-utility curve for this architecture and dataset size, and is a useful practical signal for a practitioner deciding how much privacy to purchase: budgets below roughly $\epsilon = 1$ are not simply "less accurate" but functionally non-predictive for this model.

5.2. Class-Imbalance Handling for a Small Deep Network

The finding in Section 4.1 that leaving the training set unbalanced produced the best aggregate metrics, while class-weighting and SMOTE each traded some aggregate accuracy for a modest gain in minority-class recall is worth reading carefully rather than simply as a ranking. Aggregate accuracy and F1-score are computed over a test set that is itself 69% approved, so a model that leans toward predicting approval scores well on both metrics even when it is, in effect, less discriminating about which applicants it rejects; the specificity figures in Section 4.1 make that lean visible in a way the aggregate metrics do not. Whether that trade-off is acceptable depends on the relative cost a lender places on approving a bad loan against rejecting a good one, a business decision this study does not attempt to make on the lender's behalf. What the comparison does establish is that, for a network of this size trained on 491 examples, synthetic oversampling did not provide a clear advantage over simple class-weighting. SMOTE's k-nearest-neighbour interpolation in a fifteen-dimensional engineered feature space appears, on this evidence, to introduce as much interpolation noise as it does useful synthetic signal, echoing a caution raised elsewhere in the tabular-data literature that SMOTE's benefit shrinks as feature dimensionality grows relative to dataset size.

5.3. Explaining a Privacy-Constrained Network

The convergence of permutation importance, integrated gradients and SHAP on Credit_History as the dominant predictor (Section 4.5) was computed on the non-private baseline network, and a natural further question is whether that same feature ranking survives under DP-SGD training. Because DP-SGD clips and perturbs gradients uniformly across all fifteen input features rather than targeting any one of them, there is no structural reason for the mechanism to selectively suppress the signal from Credit_History relative to weaker features in contrast to a feature-level DP mechanism, which could in principle be

tuned to protect a sensitive field like credit history more or less aggressively than others. Confirming this empirically for the DP-SGD-trained networks across the sweep, rather than only for the non-private baseline, is identified as a direct extension of this study in Section 6.

5.4. Limitations

Several limitations qualify these results. First, full-batch gradient descent was used specifically to keep privacy accounting tractable by hand; a Poisson-subsampled mini-batch implementation would benefit from privacy amplification by subsampling, on top of the Rényi-DP tightening already demonstrated in Section 4.7, at the cost of a substantially more involved accounting procedure. The gap between the two accountants in Table 7 should therefore be read as a lower bound on how much tighter a fully optimised accounting pipeline could get, not the maximum achievable tightening. Second, the primary dataset itself is small (614 records) by deep-learning standards, which limits how much a neural architecture can be expected to outperform a well-tuned ensemble model on this problem specifically, independent of any privacy mechanism (Section 5.5). Third, the external validation in Section 4.9 and 5.6 uses a single second dataset and a reduced two-point privacy sweep rather than a full independent replication, so it should be read as a directional check rather than proof of generalisation across lending contexts.

5.5. Classical Baselines and the Cost of Going Deep

Table 6 shows Random Forest numerically ahead of the non-private MLP, and histogram gradient boosting close behind it, on this 491-example training set unsurprising given the well-documented tendency for tree ensembles to match or outperform deep networks on small, low-dimensional tabular problems (Gunnarsson et al., 2021). Table 8's paired t-tests, however, do not find either gap statistically significant at five seeds, so this study does not claim the MLP is worse than

a tuned ensemble, only that neither model has a demonstrated edge at this sample size. What the deep-learning architecture offers instead of a raw accuracy advantage is the gradient-level DP-SGD mechanism itself (Section 5.1): applying an equivalent formal privacy guarantee directly to Random Forest or gradient-boosted trees is considerably less natural, since neither model is trained by gradient descent, which is precisely why the feature-level and output-level DP mechanisms reviewed in Section 2.2 remain the dominant approach for those model families. The choice between a classical ensemble and a DP-SGD-trained network is therefore better framed as a choice of privacy mechanism than as a choice of raw accuracy.

5.6. External Validation

The German Credit check in Section 4.9 was not intended to establish that this exact architecture is optimal for a second lending context: no hyperparameter re-tuning was performed for that dataset but to test whether the two qualitative claims made about the primary dataset are artefacts of that one benchmark. Both claims replicated: the non-private model reached a broadly comparable accuracy and ROC-AUC on an independently collected dataset with different features and a different population, and the DP-SGD sweep showed the same collapse-then-recover pattern across the same two noise multipliers, recovering a near-identical share of non-private accuracy (96.6% on German Credit versus 98.4% on the primary dataset) at the loosest privacy budget tested. This consistency, on a dataset the pipeline was never tuned against, is modest but genuine evidence that the privacy-utility behaviour documented in Section 4.4 reflects a property of DP-SGD applied to this style of small tabular credit-scoring problem, rather than a coincidence of the primary benchmark's specific feature distribution.

6. CONCLUSION AND FUTURE WORK

This study trained a compact multilayer perceptron for loan default prediction under an explicit, from-first-principles DP-SGD procedure, rather than applying differential privacy to the model's input features or outputs after the fact. Across a swept privacy budget from $\epsilon = 0.39$ to $\epsilon = 7.27$, the resulting network traces a smooth, monotonic privacy-utility curve and recovers 98.4% of its non-private accuracy at the most permissive budget tested, without requiring any feature-specific sensitivity analysis. Permutation importance, integrated gradients and SHAP computed directly on the trained network rather than on a separate surrogate model, independently identify credit history as the dominant predictor, consistent with the wider loan-prediction literature and providing a basic form of cross-validation between two different explanation methods. A secondary comparison of class-imbalance handling strategies found that, for this architecture and dataset size, doing nothing produced the best aggregate metrics while class-weighting and SMOTE each produced a modest, specificity-focused improvement at some cost to aggregate accuracy: a trade-off that should be resolved according to a lender's own cost structure rather than by aggregate accuracy alone.

Three directions follow directly from the limitations in Section 5.4. First, moving from full-batch to Poisson-subsampled mini-batch DP-SGD would combine the Rényi-DP tightening already demonstrated in Section 4.7 with the additional amplification-by-subsampling effect, for which full-batch training receives no benefit: a production implementation via the Opacus library (Yousefpour et al., 2021), potentially combined with training-side techniques such as tempered sigmoid activations shown elsewhere to reduce the gradient norms DP-SGD must clip (Papernot et al., 2021), would likely shift the usable part of the privacy-utility curve further toward stronger privacy than reported here. Second, applying permutation importance and integrated gradients to the DP-SGD-trained networks across the full sweep, rather than only to the non-private

baseline, would directly test whether feature attribution is preserved under privacy noise, as conjectured in Section 5.3 and as suggested by the uneven, feature-dependent degradation of explanatory power observed under privacy constraints in a related bank-failure-prediction study (Byun et al., 2025). Third, the external validation in Section 4.9 used a reduced two-point sweep on a single second dataset; a fuller replication the complete six-point sweep, cross-validated, on several independent lending datasets from different institutions or countries would more firmly establish how broadly the privacy-utility and feature-importance findings reported here generalise beyond the two benchmarks used in this study.

AUTHOR CONTRIBUTIONS

Conceptualization, M.A.S. and S.A.N.; Methodology, M.A.S. and S.A.N.; Software, M.A.S. and S.A.N.; Validation, M.A.S., S.A.N., and M.S.; Formal Analysis, M.A.S. and S.A.N.; Investigation, M.A.S., S.A.N., and M.S.; Data Curation, S.A.N.; WritingOriginal Draft Preparation, M.A.S.; WritingReview and Editing, M.A.S., S.A.N., M.S., M.U., F.S., and W.H.; Visualization, S.A.N.; Supervision, M.S.; Project Administration, M.S. All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY STATEMENT

The primary dataset analysed in this study is the publicly available Loan Prediction benchmark dataset originally released by Analytics Vidhya and mirrored on Kaggle (Section 3.1); no restrictions apply to its use for research purposes. The external validation dataset is the publicly available Statlog (German Credit Data) dataset from the UCI Machine Learning Repository (Hofmann, 1994), licensed under CC BY 4.0. All code implementing the preprocessing pipeline, the DP-SGD network, the privacy accountants, and the analyses reported in this paper, together with both datasets and a README describing how to reproduce every table and figure, is openly available at

<https://github.com/atifbhatti092/loan-dp-sgd-explainability>.

ACKNOWLEDGMENTS

The authors thank their supervisor, Mr. Muhammad Saqib, Department of Computer Sciences, University of Lahore, Sargodha Campus, for guidance throughout the design and execution of this project.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- 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>
- Abedin, M. Z., Guotai, C., Hajek, P., & Zhang, T. (2023). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems*, 9, 3559–3579. <https://doi.org/10.1007/s40747-021-00614-4>
- Azhar, N. A., Pozi, M. S. M., Din, A. M., & Jatowt, A. (2023). An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35, 6651–6672. <https://doi.org/10.1109/TKDE.2022.3179381>
- Byun, J., Lee, J., Lee, H., & Son, B. (2025). Balancing explainability and privacy in bank failure prediction: A differentially private glass-box approach. *IEEE Access*, 13. <https://doi.org/10.1109/ACCESS.2024.3523967>

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review*, 57, Article 137. <https://doi.org/10.1007/s10462-024-10759-6>
- Dwork, C. (2006). Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)* (pp. 1–12). https://doi.org/10.1007/11787006_1
- Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9, 211–407. <https://doi.org/10.1561/04000000042>
- Dwork, C., Rothblum, G. N., & Vadhan, S. (2010). Boosting and differential privacy. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science (FOCS)* (pp. 51–60). <https://doi.org/10.1109/FOCS.2010.12>
- Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Op de Beeck, M., & Snoeck, M. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295, 292–305. <https://doi.org/10.1016/j.ejor.2021.03.006>
- Hartomo, K. D., Arthur, C., & Nataliani, Y. (2025). A novel weighted loss TabTransformer integrating explainable AI for imbalanced credit risk datasets. *IEEE Access*, 13, 31045–31056. <https://doi.org/10.1109/ACCESS.2025.3541878>
- Hofmann, H. (1994). Statlog (German credit data) [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5NC77>
- Hossain, M. M., Mamun, M., Munir, A., Rahman, M. M., & Chowdhury, S. H. (2025). A secure bank loan prediction system by bridging differential privacy and explainable machine learning. *Electronics*, 14, Article 1691. <https://doi.org/10.3390/electronics14081691>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Mamun, M. A., Farjana, A., & Mamun, M. (2022). Predicting bank loan eligibility using machine learning models and comparison analysis. In *Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management (IEOM)* (pp. 12–14). Orlando, FL, USA.
- Mironov, I. (2017). Rényi differential privacy. In *Proceedings of the 2017 IEEE 30th Computer Security Foundations Symposium (CSF)* (pp. 263–275). <https://doi.org/10.1109/CSF.2017.11>
- Moradi, M., Alizadeh, M., & Ghasemi, M. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review*, 57, Article 303. <https://doi.org/10.1007/s10462-024-10854-8>
- Naresh, V. S., & Ayyappa, D. (2025). PPDNN-CRP: CKKS-FHE enabled privacy-preserving deep neural network processing for credit risk prediction. *Computational Economics*, 66, 2619–2643. <https://doi.org/10.1007/s10614-024-10731-0>

- Naresh, V. S., & Ayyappa, D. (2026). Privacy-preserving federated credit risk models: Evaluating differential privacy and homomorphic encryption techniques. *Scientific Reports*, 16, Article 4379. <https://doi.org/10.1038/s41598-025-34536-9>
- Onoufriou, N., et al. (2022). Evaluating privacy-preserving techniques in machine learning: Trade-offs and performance. *IEEE Access*, 10, 84276–84290. <https://doi.org/10.1109/ACCESS.2022.3183157>
- Papernot, N., Thakurta, A., Song, S., Chien, S., & Erlingsson, Ú. (2021). Tempered sigmoid activations for deep learning with differential privacy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10), 9312–9321. <https://doi.org/10.1609/aaai.v35i10.17124>
- Ratadiya, P., Asawa, K., & Nikhal, O. (2020). A decentralized aggregation mechanism for training deep learning models using smart contract system for bank loan prediction (arXiv:2011.10981). arXiv. <https://arxiv.org/abs/2011.10981>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
- Sarkar, T., Rakhra, M., Sharma, V., & Singh, A. (2024). An empirical comparison of machine learning techniques for bank loan approval prediction. In *Proceedings of the 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)* (pp. 137–143). <https://doi.org/10.1109/IC3SE62002.2024.10593355>
- Shinde, A., Patil, Y., Kotian, I., Shinde, A., & Gulwani, R. (2022). Loan prediction system using machine learning. *ITM Web of Conferences*, 44, Article 03019. <https://doi.org/10.1051/itmconf/20224403019>
- Spooorthi, B., Kumar, S. S., Rodrigues, A. P., Fernandes, R., & Balaji, N. (2021). Comparative analysis of bank loan defaulter prediction using machine learning techniques. In *2021 IEEE DISCOVER* (pp. 24–29). IEEE. <https://doi.org/10.1109/DISCOVER52564.2021.9663616>
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)* (pp. 3319–3328).
- Uddin, N., Ahamed, M. K. U., Uddin, M. A., Islam, M. M., Talukder, M. A., & Aryal, S. (2023). An ensemble machine learning based bank loan approval predictions system with a smart application. *International Journal of Cognitive Computing in Engineering*, 4, 327–339. <https://doi.org/10.1016/j.ijcce.2023.09.001>
- Viswanatha, V., Ramachandra, A. C., Vishwas, K. N., & Adithya, G. (2023). Prediction of loan approval in banks using machine learning approach. *International Journal of Engineering and Management Research*, 13, 7–19. <https://doi.org/10.31033/ijemr.13.4.2>
- Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., & Mironov, I. (2021). Opacus: User-friendly differential privacy library in PyTorch (arXiv:2109.12298). arXiv. <https://arxiv.org/abs/2109.12298>
- Zheng, Y., Wu, Z., Yuan, Y., Chen, T., & Wang, Z. (2020). PCAL: A privacy-preserving intelligent credit risk modeling framework based on adversarial learning (arXiv:2010.02529). arXiv. <https://arxiv.org/abs/2010.02529>