Siamese Network for Microfinance Credit Scoring under Low-Data Regimes

Bogomazov Vladislav
Bocconi Students for Machine learning
Bocconi University
Email: vladislav.bogomazov@studbocconi.it

***Abstract*—Credit scoring in microfinance is challenging due to limited data and highly imbalanced repayment patterns. In this study, we adopt a Siamese network approach to predict default and compare its performance with Random Forest, CatBoost, Gradient Boosting Decision Trees, and Gaussian Process Classifier on a subset of the *Kiva Crowdfunding for Good* dataset. Models were evaluated under both balanced (50:50) and imbalanced (95:5) conditions using metrics including Accuracy, F1, ROC AUC, PR AUC, Log Loss, Brier Score, and KS Statistic. Under balanced splits, Random Forest achieved the highest overall performance (ROC AUC 0.90, PR AUC 0.90), while its performance degraded significantly under extreme imbalance. The Siamese network (representation-based) and Gaussian Process Classifier (probabilistic) maintained robustness, achieving the highest PR AUC in imbalanced settings (0.81 and 0.82, respectively). These results highlight the potential usefulness of Siamese networks for credit scoring in highly imbalanced microfinance datasets.**

***Index Terms*—Microfinance, Credit Scoring, Low-Data Regime, Representation Learning, Siamese Network, Machine Learning**

I. Introduction

Global poverty and financial exclusion persist as major barriers to economic development. Nearly 700 million people—8.5% of the global population—live in extreme poverty, subsisting on less than $2.15 per day, and roughly 3.5 billion people (44%) fall below a $6.85 daily threshold [1]. These populations are disproportionately concentrated in climate-vulnerable regions, with over 80% of the 1.4 billion unbanked adults facing heightened exposure to economic and environmental shocks [4]. Limited access to formal financial services exacerbates vulnerability, particularly for populations with scarce credit histories.

Microfinance institutions (MFIs) have attempted to address this gap by providing small loans, savings, and insurance products [21]. However, MFIs face persistent operational constraints, including high transaction costs, limited sustainability, and unreliable borrower creditworthiness [29]. Traditional credit scoring methods frequently fail for clients with sparse financial histories, resulting in elevated default rates and restricted coverage. While recent digital initiatives—such as MyBucks Malawi Limited—demonstrate the potential of data-driven approaches [6] [10], existing machine learning (ML) methods largely assume abundant, structured data, limiting applicability in low-data microfinance contexts [22].

This study addresses these limitations by proposing a credit scoring framework specifically tailored for small-scale MFIs operating under data scarcity and class imbalance. By leveraging representation learning techniques, including Siamese networks and transformer-based encoders, our framework integrates alternative data sources to enhance predictive accuracy and operational efficiency. Unlike conventional approaches, this framework is designed to remain robust when borrower data are sparse or heterogeneous, directly targeting the critical challenges that have hindered ML adoption in microfinance.

II. Related Work

A. Traditional Credit Scoring in Microfinance

Microfinance institutions (MFIs) have historically faced unique challenges in credit scoring compared to conventional banks. Most relied on the experience and intuition of credit officers rather than formal scoring systems, and when scoring systems existed, they typically used only basic financial indicators [5]. Schreiner [5] categorizes microfinance credit scoring into statistical approaches, which estimate default risk from historical loan performance, and judgmental approaches, which rely on credit officer expertise. The latter has been dominant, reflecting the scarcity of reliable historical data. Empirical studies indicate that institutions adopting formal credit scoring generally outperform those relying solely on manual assessments [19].However, traditional models systematically exclude borrowers with limited or no credit histories—the very populations microfinance aims to serve [28]. This limitation highlights the need for alternative methods capable of evaluating creditworthiness with sparse or unconventional data.

B. Machine Learning for Microfinance Credit Scoring

Recent research demonstrates that machine learning (ML) techniques can substantially improve microfinance credit scoring, especially for borrowers with limited credit histories. Comparative studies consistently show that ML models outperform conventional approaches, with ensemble methods such as Random Forests and Gradient Boosting achieving some of the highest predictive accuracy [23], [24]. Neural networks and boosting algorithms have produced improvements in AUC exceeding 15 percent over traditional statistical models, while hybrid approaches combining multiple models often yield the strongest overall performance. Despite these advances, most ML applications assume relatively large, well-structured datasets—conditions rarely met in typical MFIs. For instance, a study employing contrastive learning and domain adaptation

achieved an AUC of 0.714 and profits of 458,686, but it used a dataset of over 311,000 loan applications [16]. Most MFIs operate with far smaller datasets, underscoring the need for approaches that perform reliably in low-data regimes.

C. Alternative and Emerging Approaches

Alternative data and new methods are improving microfinance credit scoring. Mobile usage, digital payments, psychometrics, social media, and utility records can raise accuracy by up to 12%, with mobile data alone reaching 89%. Psychometric tools assess traits like integrity and risk attitudes, expanding access for those without formal histories. Evidence from Ethiopia and Peru shows greater inclusion—especially for women and SMEs—without harming repayment, though results vary by context. [17]

Incorporating social and environmental factors aligns scoring with MFI missions, but tools like Social NPV and MEPI face standardization challenges. Group lending sustains repayment rates above 90%, [30] though MFIs often shift to individual loans for established clients. [] Digital and mobile scoring is growing rapidly, lowering costs and enabling instant approvals, yet field studies in Kenya and Tanzania report default rates of 47–56%, underscoring transparency and adoption issues. [31]Overall, these innovations are context-dependent, requiring adaptation and supportive regulation. [7] [25] [26]

D. Research Gap

Despite these advances, most existing approaches either rely on large datasets or are context-specific, limiting their generalizability in typical MFI settings. There is a clear need for methods that utilise the predictive power of machine learning , while remaining robust in low-data environments. This work addresses this gap by developing a scalable credit scoring framework tailored for small-scale MFIs, leveraging machine learning to improve accuracy and financial inclusion.

III. Siamese Network Framework

A. Motivation

Conventional classifiers in credit scoring rely on absolute class boundaries and require large labeled datasets. In contrast, our setting involves a deliberately constrained sample (1,000 loans) to simulate a low-data regime. To address this, we adopt a Siamese network architecture, originally proposed for metric learning [8], which learns a representation space by modeling pairwise similarity.

A key advantage of this approach in few-shot learning (FSL) contexts is its ability to mitigate overfitting. Instead of training on N raw examples, a Siamese network constructs $\frac{N(N-1)}{2}$ pairs by comparing every sample with every other sample, effectively augmenting the dataset [27]. This combinatorial pairing greatly increases the number of training signals and reduces the risk of overfitting under limited data. At the same time, the model functions as a representation learning mechanism: it extracts discriminative low-dimensional features while discarding redundant information in the input data [27]. These properties make Siamese networks particularly well-suited for low-resource credit scoring.

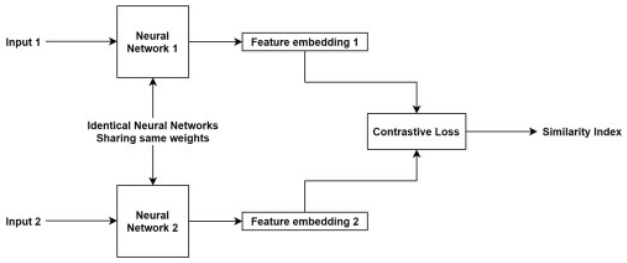

Fig. 1: Architecture of the Siamese network: two identical encoders map loan records into embeddings, followed by distance computation and contrastive loss training.

B. Architecture and Objective

The Siamese network consists of two identical encoder branches, $f_\theta(\cdot)$, sharing parameters θ. Given two loan records x_i and x_j, the encoders produce embeddings:

$$h_i = f_\theta(x_i), \quad h_j = f_\theta(x_j).$$

Weight sharing ensures that both inputs are mapped into a consistent feature space. To quantify similarity, we use Euclidean distance:

$$D(h_i, h_j) = \|h_i - h_j\|_2.$$

Pairs are constructed such that $y = 1$ if two loans share the same repayment class (e.g., both "regular"), and $y = 0$ otherwise.

In practice, these pairs are generated by randomly selecting combinations of loans from the dataset. For each anchor loan, a partner is chosen either from the same repayment class to form a positive pair or from a different class to form a negative pair. This random pair sampling ensures a diverse set of relationships for the network to learn from, while still covering all possible pairwise interactions in the $\frac{N(N-1)}{2}$ combinatorial space. By converting a limited set of individual loans into a much larger set of training pairs, the model effectively amplifies the amount of informative training signals, which is particularly beneficial in low-data environments. This approach allows the Siamese network to learn more robust and generalizable embeddings, capturing subtle patterns of similarity and dissimilarity across borrowers.

The network is trained with a contrastive loss [**?**]:

$$\mathcal{L} = y \cdot D(h_i, h_j)^2 + (1 - y) \cdot \max\left(0, m - D(h_i, h_j)\right)^2,$$

where m is a margin hyperparameter. This objective enforces compact clusters for loans with the same repayment outcome while maximizing separation between dissimilar ones.

A schematic overview of the Siamese network used in this study is presented in Fig. 1. The diagram illustrates the dual-branch structure, weight sharing, embedding generation, and distance computation between loan pairs.

C. Embedding Utilization

After pretraining the encoder, we freeze its parameters and use the learned embeddings for downstream classification. K-Nearest Neighbors (distance-based), was chosen as the main classifier.

D. Evaluation Protocol and Statistical Testing

Performance was assessed on the held-out test set using widely adopted credit scoring metrics. Accuracy is reported for completeness, but greater emphasis is placed on metrics that better capture discriminative ability (ROC AUC, PR AUC), calibration (Log Loss, Brier Score), and class separation (Kolmogorov–Smirnov statistic), which are standard in financial risk modeling [**?**].

All models were evaluated through a unified evaluation framework. This framework encompassed assessment of embedding quality via KNN classification, computation of the standard performance metrics listed above, bootstrap resampling ($n = 1000$) to estimate 95% confidence intervals for each metric, and confusion matrix analysis to characterize classification errors. The same evaluation methodology was applied consistently across all experimental conditions, including both the balanced 50:50 and imbalanced 95:5 data splits.

E. Rationale and Limitations

This design directly addresses the challenges of low-data credit scoring: (i) pairwise training increases the number of effective training examples, mitigating overfitting, (ii) contrastive loss encourages generalization across borrowers with heterogeneous features.

IV. Experiments and Results

A. Experimental Setup

We used the *Kiva Crowdfunding for Good* dataset from Kaggle, which contains 671,205 loan records across 19 attributes. To simulate a low-data regime typical in microfinance, 1,000 records were sampled and split into training (640), validation (160), and testing (200) sets.

Columns directly tied to loan funding (e.g., `id`, `funded_amount`, `loan_amount`, `currency`, `partner_id`, and timestamps) were excluded to avoid leakage. The target variable was binarized: regular repayments (weekly or monthly) were encoded as 0, and non-regular (bullet or irregular) as 1. Retained predictors include loan activity, sector, country, term length, borrower count, borrower gender, and loan issue date. Borrower gender was encoded as: 0 (male-only), 1 (female-only), 2 (unknown), and 3 (mixed).

High-cardinality categorical features (`activity`, `sector`, `country`) were represented using ordered target encoding, while numerical features were standardized and missing values imputed. (replaced with estimated values) Figure 2 illustrates the sectoral distribution of loans, showing the dominance of Agriculture, Food, and Retail.

Two evaluation scenarios were considered: (i) a balanced 50:50 split between repaid and defaulted loans, and (ii) an imbalanced 95:5 split reflecting real-world repayment distributions. Importantly, all models were tested and validated on the same dataset constructed from the 50:50 sampling to ensure comparability across approaches.

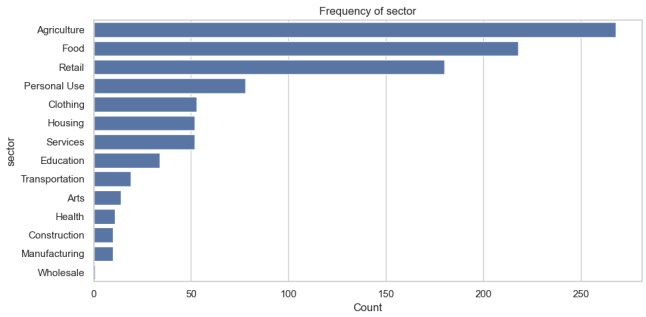

Fig. 2: Distribution of loans across economic sectors in the sampled dataset.

B. Model Performance

Table I reports results under the balanced 50:50 split. Random Forest achieved the strongest performance across nearly all metrics, with CatBoost and GBDT close behind. Representation-based method Siamese+KNN lagged slightly but remained competitive.

[t]

However, microfinance repayment data is rarely balanced. To reflect the real-world distribution, we constructed a 95:5 dataset. Table II summarizes performance under this extreme imbalance. Unlike the balanced setting, traditional ensembles degraded severely (Random Forest $\sim$56.5%, CatBoost $\sim$58%), while representation-based model demonstrated robustness (Siamese+KNN at 72.0%)

Figure 3 illustrates this trend across a continuum of splits from 75:25 to 95:5, showing that Siamese degrades more gracefully than ensemble methods. Splits from 75:25 to 95:5 are shown because the trend is symmetrical, so this range sufficiently illustrates how Siamese degrades compared to ensembles without redundancy

The Receiver Operating Characteristic Area Under the Curve (ROC AUC) is a scalar metric that summarizes the performance of a binary classifier across all possible decision thresholds. It is defined as the probability that the classifier assigns a higher score to a randomly chosen positive instance than to a randomly chosen negative instance. Geometrically, it corresponds to the area under the ROC curve, which plots the true positive rate (sensitivity) against the false positive rate (1 – specificity). A value of 0.5 indicates performance no better than random guessing, while a value of 1.0 represents perfect discrimination between classes.

Results for 50:50 split can be seen on the graph 4 The ROC AUC results highlight Random Forest as the strongest performer (0.90), with CatBoost (0.87) and GPC (0.84) following closely behind. Multilayer GBDT achieved a moderate score of 0.82, while the Siamese Network + KNN lagged slightly

Model	Accuracy	F1	ROC AUC	PR AUC	LogLoss	Brier	KS
Random Forest	**0.8144**	**0.8148**	**0.8992**	**0.8991**	**0.4060**	**0.1302**	**0.6721**
CatBoost	0.8001	0.7975	0.8739	0.8633	0.4483	0.1421	0.6321
GBDT	0.7949	0.7939	0.8211	0.7893	1.8102	0.2012	0.6389
Siamese+KNN	0.7750	0.7783	0.8300	0.7718	3.8410	0.1852	0.5528
GPC	0.7500	0.7475	0.8447	0.7959	0.4982	0.1628	0.5180

TABLE I: Model performance under balanced conditions (50:50 split).

Model	Accuracy	F1	ROC AUC	PR AUC	LogLoss	Brier	KS
Siamese + KNN	**0.7200**	**0.7705**	0.7377	0.6498	8.7791	**0.2786**	0.4784
Random Forest	0.5655	0.6922	**0.7940**	**0.7505**	2.9202	0.3439	**0.5280**
CatBoost	0.5813	0.7118	0.6515	0.5904	**1.3402**	0.3024	0.4901
GPC	0.5655	0.6922	**0.7940**	**0.7505**	2.9202	0.3439	**0.5280**
Multilayered	0.6250	0.7148	0.6316	0.5664	4.8203	0.3750	0.2631

TABLE II: Performance metrics of different models under unbalanced conditions (95:5 split).

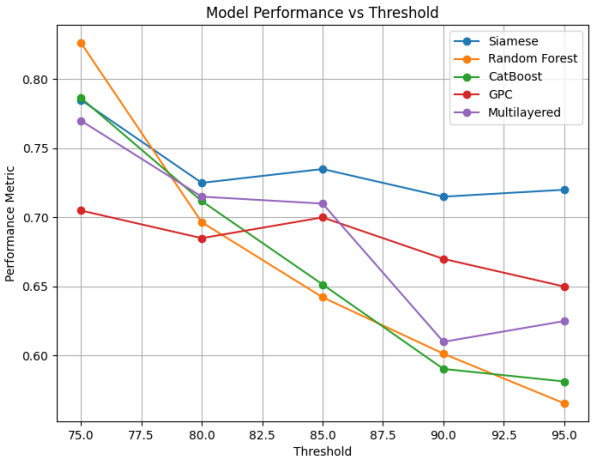

Fig. 3: Best performing models' accuracies over different splits from 75:25 to 95:5.

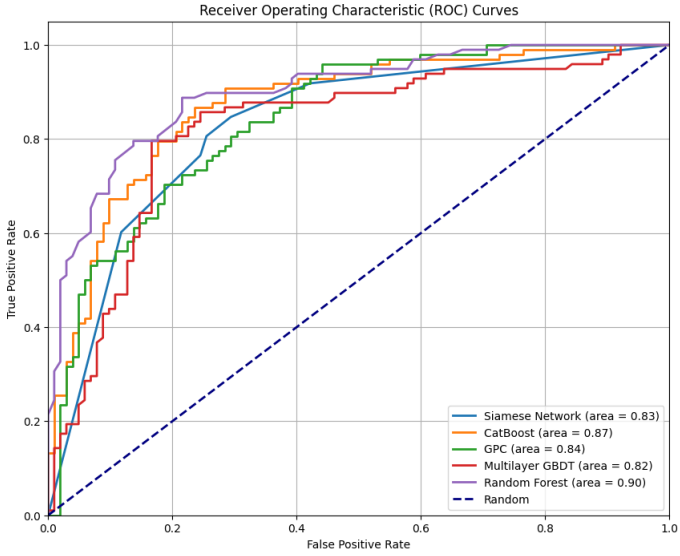

Fig. 4: ROC curves for selected models under balanced 50:50 split.

at 0.83. These results indicate that ensemble-based models, particularly Random Forest, are most effective at ranking positive instances higher than negatives across thresholds. GPC also performs well in terms of discrimination, though its lower accuracy and F1 score suggest trade-offs in classification at fixed thresholds. The Siamese Network's ROC AUC is comparable to Multilayer GBDT, but given its much poorer calibration and high log loss, the quality of its probabilistic ranking is less reliable. Overall, the ROC AUC analysis reinforces the strong discriminative ability of Random Forest and CatBoost in this dataset.

To better understand, ROC graph can be seen on 5. Based on the ROC AUC results, the Gaussian Process Classifier (0.83) achieved the highest discriminative performance, reflecting its strength in modeling complex non-linear relationships and producing well-calibrated probabilities, as also supported by our calibration curve analysis. Random Forest (0.79) and CatBoost (0.77) followed closely, consistent with the strong performance typically observed in tree-based ensemble methods, which are well-suited for capturing feature interactions and handling heterogeneous data. The Siamese Network combined with KNN (0.72) demonstrated moderate performance, indicating that while the learned embeddings provided some separation—as seen in the t-SNE visualization—the simplicity of the KNN classifier limited its effectiveness compared to more sophisticated ensemble or probabilistic models. Finally, the Multilayer GBDT (0.63), despite the corrected data, underperformed significantly, suggesting potential challenges with its architectural complexity or hyperparameter tuning. Overall, these results highlight the relative strengths of Gaussian Processes and tree-based methods for this dataset, while also illustrating the representational promise—but practical limitations—of embedding-based approaches like Siamese networks.

However, for imbalanced Datasets, usually a more reflective characteristic would be PR AUC. While ROC AUC evaluates overall discrimination, PR AUC focuses on the minority class, which is critical in imbalanced microfinance datasets.

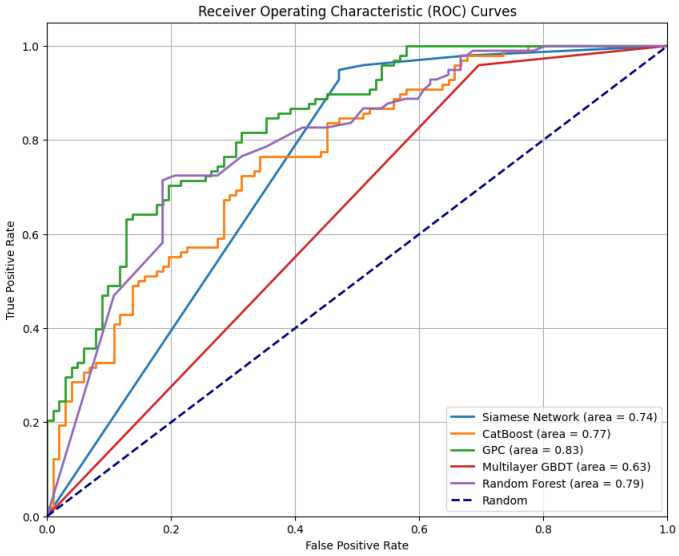

Fig. 5: ROC curves for selected models under unbalanced 95:5 split.

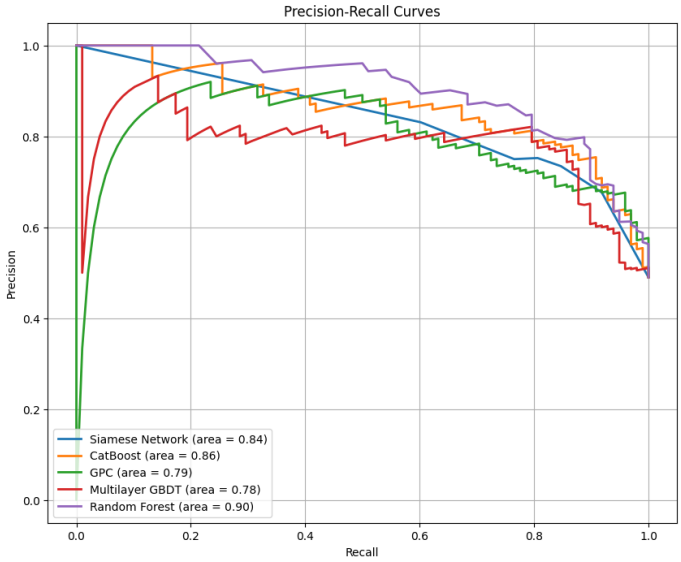

Fig. 6: Precision-Recall curves for selected models under balanced 50:50 split.

The Precision–Recall Area Under the Curve (PR AUC) is a performance metric that evaluates the trade-off between precision (the proportion of predicted positives that are truly positive) and recall (the proportion of actual positives that are correctly identified) across all decision thresholds. It is computed as the area under the precision–recall curve, which is particularly informative for imbalanced datasets where the positive class is rare. Unlike ROC AUC, which considers both classes equally, PR AUC focuses on the classifier's ability to correctly identify the minority class, making it especially relevant in applications where false negatives carry higher costs. A higher PR AUC indicates better performance in balancing precision and recall.

Balanced Split Precision-recall curve can be seen on 6

In terms of PR AUC, which emphasizes performance on the minority class, Random Forest again leads with a value of 0.90, followed by CatBoost (0.86) and GPC (0.79). Multilayer GBDT achieves 0.78, while the Siamese Network + KNN scores 0.77. These results underscore that Random Forest and CatBoost not only discriminate well overall, but also maintain strong precision-recall trade-offs in the imbalanced setting of this dataset. GPC, while solid in ROC AUC, shows a larger drop in PR AUC, reflecting lower effectiveness in correctly identifying positive instances without excessive false positives. The Siamese Network achieves only moderate precision-recall performance, consistent with the observed overlap in its t-SNE embeddings, which makes KNN less effective for separating classes. Thus, Random Forest and CatBoost stand out as the most effective models when the minority class is prioritized.

PR AUC for an unbalanced split can be seen on Fig. 7

Based on the PR AUC results, the Gaussian Process Classifier (0.82) achieved the highest performance, with the Siamese Network + KNN (0.81) following very closely. Since PR AUC emphasizes precision-recall trade-offs for the minority class,

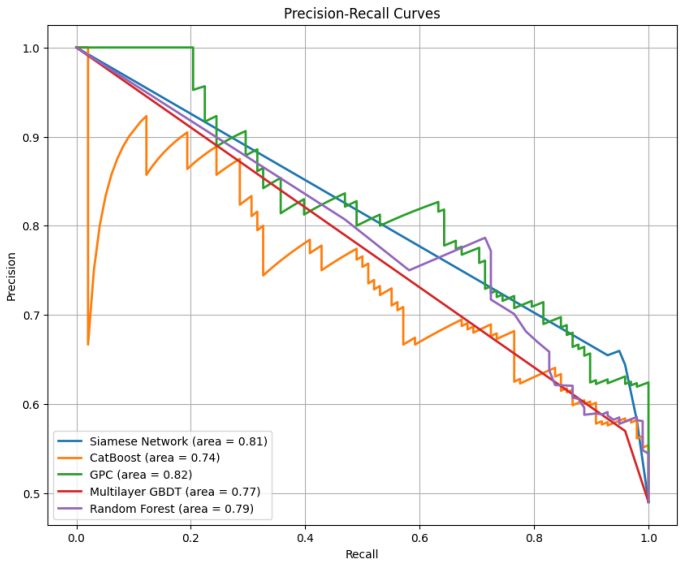

Fig. 7: Precision-Recall curves for selected models under unbalanced 95:5 split.

these results suggest that GPC's ability to capture complex non-linear relationships and model uncertainty provides it with an advantage in imbalanced settings, while the Siamese Network's embedding approach effectively grouped minority-class instances in a way that allowed KNN to achieve a strong balance between precision and recall. Random Forest (0.79) and Multilayer GBDT (0.77) performed moderately, indicating that while tree-based models are generally robust to imbalance, they were less effective than GPC or the Siamese Network in optimizing the precision-recall trade-off on this dataset. CatBoost (0.74) showed the weakest performance, suggesting that without further class-weight tuning, its boosting procedure

struggled to prioritize minority-class recall. Importantly, confidence interval analysis highlights that GPC's performance is likely statistically significantly higher than that of the Siamese Network and Multilayer GBDT, while differences among GPC, Random Forest, and CatBoost are not statistically significant due to overlapping intervals. Overall, the results indicate that GPC and the Siamese Network stand out as the strongest models when evaluated on PR AUC, underscoring the value of both probabilistic modeling and representation learning approaches in highly imbalanced microfinance datasets.

C. Calibration

A calibration curve evaluates the reliability of predicted probabilities from a classifier. It plots the predicted probability of the positive class on the x-axis against the observed frequency of positives on the y-axis, typically after binning predictions into intervals. A perfectly calibrated model produces a curve that follows the diagonal line, meaning that predictions correspond closely to true probabilities (e.g., predictions of 0.8 correspond to 80% actual positives). Deviations above the diagonal indicate underestimation, while deviations below indicate overestimation. Calibration is often quantified using metrics such as Brier score or log loss, which measure the mean squared error or likelihood discrepancy between predicted probabilities and actual outcomes. Well-calibrated models are particularly important when predicted probabilities are used for decision-making or risk assessment.

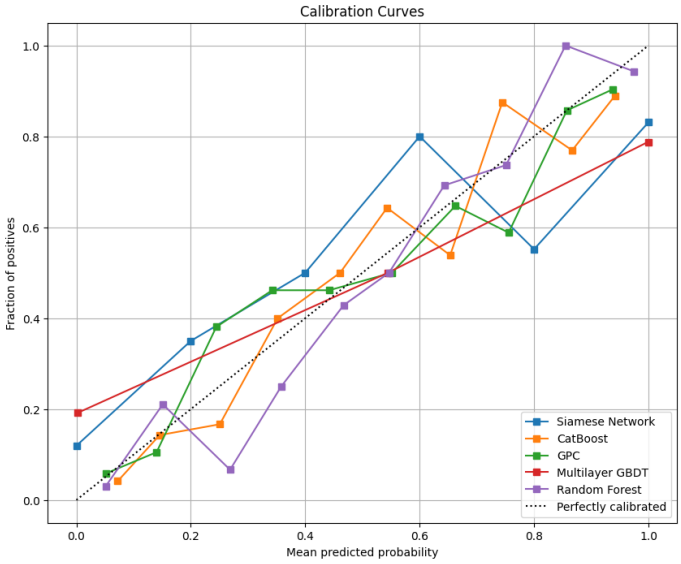

Fig. 8: Precision-Recall curves for selected models under balanced 50:50 split.

For a balanced dataset, calibration curves (Figure 8) reveal further distinctions between the models. Random Forest and CatBoost are both relatively well-calibrated, with predicted probabilities aligning closely to the diagonal, suggesting that their outputs can be interpreted as reliable confidence scores. GPC is also reasonably calibrated, consistent with its Bayesian framework. By contrast, Multilayer GBDT and the Siamese Network deviate significantly from the diagonal. In particular, the Siamese Network + KNN shows the poorest calibration, with its probabilities systematically misaligned with observed frequencies, reflected also in its very high log loss (3.8410) and elevated Brier score. This indicates that while the Siamese embeddings capture some separation in feature space, the probabilities derived from KNN are not reliable indicators of true class likelihood. Overall, Random Forest and CatBoost provide the best balance of discriminative power and calibrated probabilities for this classification task.

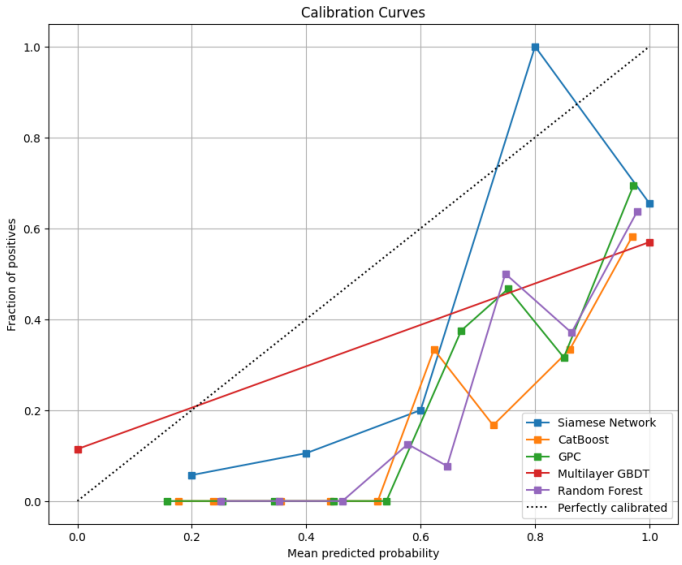

Fig. 9: Calibration curve for 95:5 imbalanced dataset.

For 95:5 split, the calibration analysis (Figure 9) highlights clear differences across models in the reliability of their predicted probabilities. The Gaussian Process Classifier (GPC) shows the strongest calibration, with its curve staying closest to the diagonal across most of the probability range. This indicates that its probabilistic outputs are highly trustworthy—for example, when the GPC predicts a probability of 0.8, approximately 80% of such cases are indeed positive. Interestingly, despite its relatively poor AUC scores, the Multilayer GBDT demonstrates comparatively better calibration in certain ranges, outperforming the Siamese Network, CatBoost, and Random Forest in terms of alignment with the diagonal. Among the tree-based ensembles, Random Forest tends to overestimate probabilities in the mid-range, while CatBoost shows a mix of over- and underestimation across different ranges, both exhibiting typical calibration deviations observed in tree-based models without post-hoc correction. The Siamese Network + KNN performs worst in this regard: its curve lies above the diagonal at low probability ranges (underestimation) and below it at high ranges (overestimation), reflecting the limitations of translating learned embedding distances into well-calibrated probabilities via a simple KNN. These results are consistent with the known strengths of Gaussian Processes in producing well-calibrated probabilistic predictions, the mixed calibration behavior of tree ensembles, and the inherent chal-

lenges of probability estimation in embedding–KNN pipelines. Overall, the GPC emerges as the most reliable model for applications where calibrated probabilities are critical.

D. Confusion Matrix

Confusion matrices (fig. 10 and 11) provide insight into error distributions. Under 50:50 balance, CatBoost and DeepGBM maintained low misclassifications, while Random Forest showed a tendency toward false negatives. The Siamese+KNN model produced more false positives but preserved minority recall.

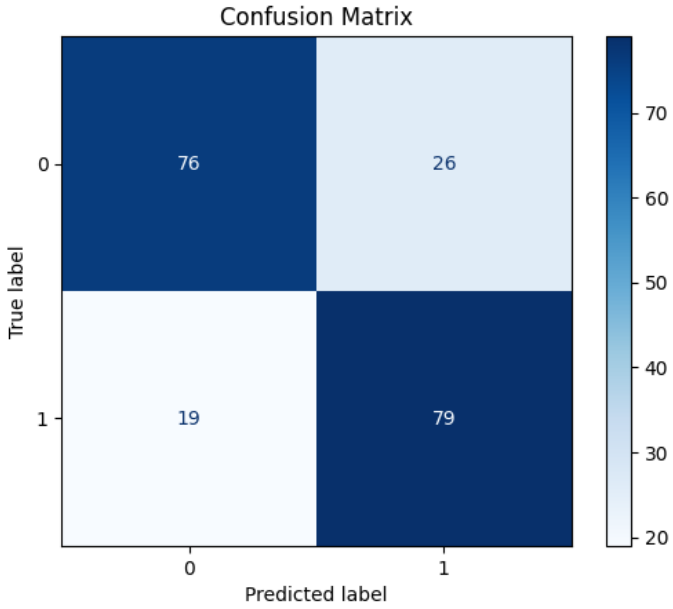

Fig. 10: Siamese network's confusion matrix under 50:50 split.

In contrast, under 95:5 imbalance, confusion matrices show ensembles collapsing to trivial majority-class predictions, while Siamesecorrectly identified non-trivial fractions of minority cases of 94/98.

E. Embedding Analysis

For a 50:50 split (Figure [12]), the average intra-class distance was 0.44, while the average inter-class distance was 1.46, meaning same-class embeddings were over three times closer than embeddings from different classes. Under a 95:5 split (Figure [13]), this separation became even more pronounced: intra-class distance dropped to 0.19 and inter-class distance increased to 1.80, yielding a nearly 9:1 ratio. These results indicate that the Siamese encoder forms a highly discriminative embedding space, where minority and majority classes remain tightly clustered and well-separated even under extreme imbalance, which helps explain its superior performance over traditional GBDT models in microfinance data.

F. Ablation Studies

V. Ablation Studies

To understand the contribution of individual design choices in our framework, we conducted a series of ablation studies.

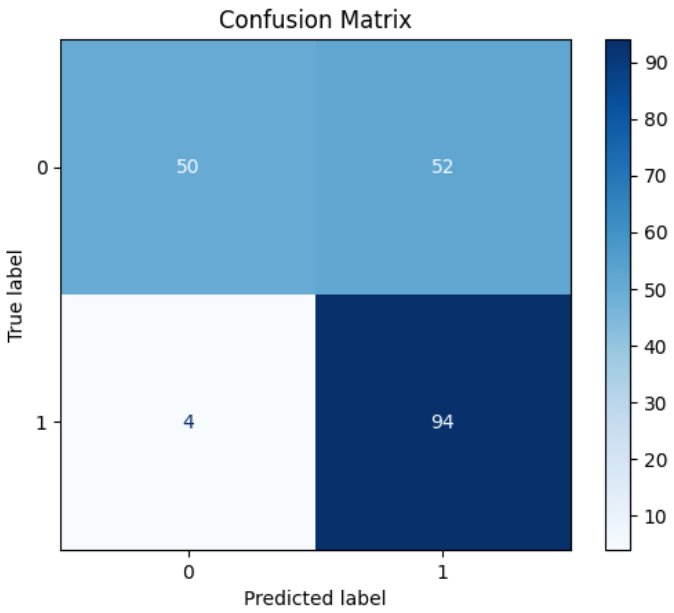

Fig. 11: Siamese network's confusion matrix under 95:5 split.
.

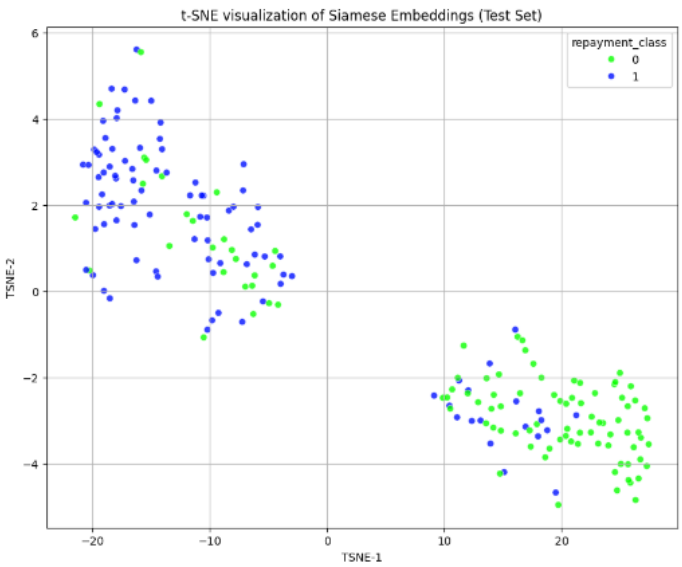

Fig. 12: t-SNE visualization of Siamese embeddings under 50:50 split.

Each variant isolates a specific component or strategy, allowing us to quantify its impact on predictive performance and probability calibration. Metrics reported include ROC AUC, PR AUC, F1 Score, Accuracy, Log Loss, Brier Score, and KS Statistic, complemented by ROC, Precision-Recall, and calibration curves for visual interpretation.

A. Ablation Overview

Each ablation serves a distinct purpose:

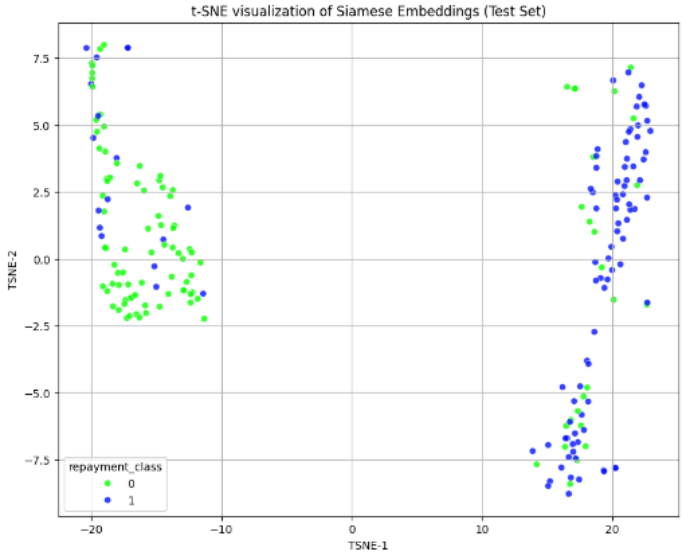

Fig. 13: t-SNE visualization of Siamese embeddings under 95:5 split. Stronger inter-class separation emerges despite imbalance.

- **Baseline vs. Removing Siamese:** Evaluates whether the encoder provides added value beyond raw features with Logistic Regression.
- **Feature Representation:** Tests how encoding strategies (scaled, one-hot, target encoding) affect model discrimination and calibration.
- **Low-Data Regime:** Assesses robustness when only a fraction of the training data is available (10% in our experiments).
- **Data Augmentation:** Gaussian noise and feature masking test whether synthetic variability improves generalization and ranking performance.
- **Loan Feature Ablation:** Examines the importance of domain-specific features for repayment prediction.
- **Calibration:** Platt scaling, isotonic scaling, and their combination evaluate how probability calibration affects both ranking metrics and threshold-dependent decisions.
- **Alternative Classifier:** Replacing Logistic Regression with KNN tests the sensitivity of downstream classification to the choice of model on Siamese embeddings.

B. Results and Interpretation for Balanced split

Full results for the Balanced split dataset are shown in the table III

a) Baseline vs. Removing Siamese Encoder: Removing the Siamese encoder leaves ROC AUC nearly unchanged (0.853) but decreases F1 from 0.783 to 0.750. This suggests that while the encoder does not significantly affect ranking metrics, it improves threshold-based classification, particularly for minority classes, as supported by the calibration curves in Fig. **??**.

b) Feature Representation: Target encoding achieves the highest ROC AUC (0.869) and PR AUC (0.851), followed closely by one-hot encoding. Scaled features perform slightly worse. The ROC and PR curves in Fig. 14–15 illustrate that feature representation strongly affects discrimination, highlighting the importance of encoding strategies in tabular microfinance data.

c) Low-Data Regime: Using only 10% of the training data reduces ROC AUC (0.846) and PR AUC (0.768), but F1 reaches the highest observed value (0.820). This demonstrates that the Siamese encoder improves robustness in data-scarce settings, likely by learning embeddings that preserve class separation even with limited samples.

d) Data Augmentation: Gaussian noise and feature masking increase ROC/PR AUC (up to 0.860/0.858) but slightly reduce F1. This indicates that augmentation enhances ranking performance but introduces variability in threshold-based classification. Figures 14 and 15 visualize these improvements.

e) Loan Feature Ablation: Removing loan-related attributes severely degrades ROC AUC (0.824) and PR AUC (0.731), while F1 and accuracy remain moderate. This confirms that domain-specific features are essential for reliable repayment prediction.

f) Calibration: Calibration substantially affects the balance between ranking and threshold metrics. Platt scaling improves F1 (0.798) and accuracy (0.800) but slightly reduces ROC/PR AUC. Isotonic calibration improves ROC/PR AUC (0.866/0.859) but maintains moderate F1. Combining isotonic with Platt achieves the highest ROC AUC (0.874) and lowest Brier score (0.141), demonstrating that calibration can be tuned to optimize for different evaluation criteria. Calibration curves in Fig. **??** confirm these trends quantitatively.

g) Alternative Classifier: Replacing Logistic Regression with KNN (k=10) on Siamese embeddings results in weaker ROC/PR AUC (0.838/0.777) and extremely poor calibration (Log Loss = 3.156). This indicates that Logistic Regression is a more suitable downstream classifier and that Siamese embeddings are not universally robust across all classifiers without tuning.

C. Results and Interpretation for Unbalanced split

Full results for the Balanced split dataset are shown in table IV

a) Baseline vs. Siamese Encoder.: Removing the Siamese encoder while using Logistic Regression on scaled features increases ROC AUC from 0.757 to 0.844 but lowers F1 from 0.760 to 0.713, indicating that the encoder mainly improves threshold-based classification, particularly for minority classes (see calibration curves in Fig. **??**).

b) Feature Representation.: One-hot encoding improves ROC AUC (0.807) and F1 (0.752) over scaled features, while target encoding slightly reduces ROC AUC (0.727) but maintains similar F1 (0.739), highlighting the importance of encoding strategies in tabular microfinance data (Fig. **??**).

TABLE III: Ablation Study Results for Siamese Network for balanced datset

Experiment	ROC AUC	PR AUC	F1 Score	Accuracy	Log Loss	Brier Score	KS Statistic
Baseline - Siamese+LR (Scaled Features)	0.853	0.825	0.783	0.775	0.557	0.178	0.590
No Siamese (LR on Scaled Features)	0.853	0.837	0.750	0.760	0.476	0.159	0.560
Siamese+LR (OHE Features)	0.858	0.844	0.780	0.775	0.546	0.176	0.573
Siamese+LR (Target Encoded Features)	0.869	0.851	0.777	0.770	0.544	0.178	0.577
Siamese+LR (Low Data 10%)	0.846	0.768	0.820	0.820	0.531	0.160	0.663
Gaussian Noise Augmentation	0.860	0.858	0.765	0.760	0.578	0.184	0.559
Feature Masking Augmentation	0.858	0.842	0.775	0.770	0.555	0.178	0.563
Ablate Loan Features	0.824	0.731	0.788	0.790	0.867	0.200	0.612
Platt Calibration	0.849	0.780	0.798	0.800	0.637	0.178	0.610
Platt + Isotonic Calibration	0.849	0.780	0.798	0.800	0.502	0.161	0.610
Isotonic Calibration	0.866	0.859	0.773	0.765	0.547	0.179	0.556
Isotonic + Platt Calibration	0.874	0.847	0.753	0.780	0.575	0.141	0.556
Siamese+KNN (k=10) (Scaled Features)	0.838	0.777	0.778	0.775	3.156	0.174	0.583

c) Low-Data Regime.: Using only 10% of the training data with the Siamese encoder yields ROC AUC 0.829, PR AUC 0.817, and F1 0.766, demonstrating robustness in data-scarce settings by preserving class separation

d) Data Augmentation.: Gaussian noise and feature masking slightly increase F1 (0.752–0.765) but do not improve ROC/PR AUC over the baseline, indicating that augmentation stabilizes threshold-based metrics without enhancing ranking performance (Figs. **??** and **??**).

e) Loan Feature Ablation.: Removing loan-specific attributes reduces F1 to 0.658 and ROC AUC to 0.768, confirming the importance of domain-specific features for reliable repayment prediction

f) Calibration.: Platt and isotonic calibration improve F1 and Brier score, with isotonic + Platt achieving ROC AUC 0.832 and Brier 0.162, demonstrating that calibration can be tuned to optimize ranking versus threshold-based metrics (Fig. **??**, Table **??**).

g) Alternative Classifier.: Using KNN (k=10) on Siamese embeddings results in ROC AUC 0.738, PR AUC 0.650, F1 0.771, and very poor calibration (Log Loss 8.779), highlighting the importance of downstream classifier choice

D. Summary

Overall, the ablation studies reveal that:

- Feature encoding and calibration have the largest effect on ranking metrics.
- The Siamese encoder provides tangible improvements in threshold-based classification, particularly in low-data regimes.
- Data augmentation enhances ROC/PR performance but has modest effects on F1.
- Domain-specific loan features are indispensable for reliable predictions.
- The choice of downstream classifier significantly affects both ranking and calibration.

These insights support the design choices in our framework and demonstrate the robustness of our approach under various perturbations and data regimes.

VI. Future Work

While our study demonstrates the potential of representation learning and ensemble methods for credit scoring in low-data, imbalanced microfinance settings, several avenues remain for further exploration.

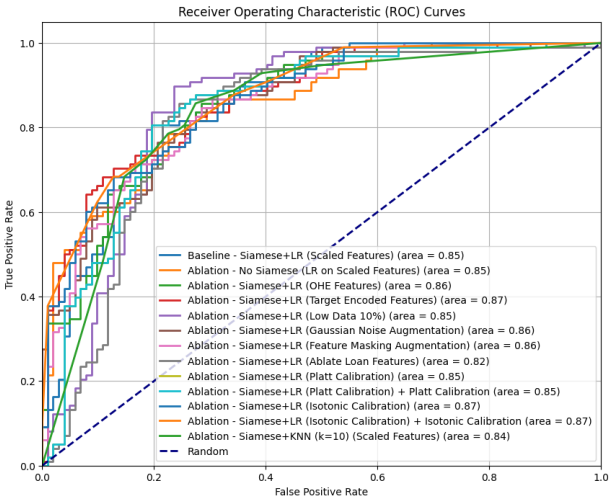

Fig. 14: ROC curves for ablation study models.

A. calibration

First, advanced calibration techniques could be integrated to improve the reliability of probabilistic predictions, particularly for embedding-based models like the Siamese Network, which exhibited high log loss despite reasonable ROC AUC scores. Approaches such as temperature scaling, isotonic regression, or Bayesian post-processing could enhance confidence estimates and reduce misclassification risk in high-stakes financial applications.

B. Temporal modeling

Third, our evaluation primarily considered static models trained on snapshot datasets. Temporal modeling of borrower behavior, such as recurrent or transformer-based architectures, could capture evolving repayment patterns, potentially improving prediction of irregular or bullet repayments over time and further optimizing precision-recall trade-offs in minority classes.

TABLE IV: Ablation Study Results for Unbalanced dataset

Experiment	ROC AUC	PR AUC	F1 Score	Accuracy	Log Loss	Brier Score	KS Statistic
Baseline - Siamese+LR (Scaled Features)	0.757	0.631	0.760	0.710	1.432	0.289	0.521
No Siamese (LR on Scaled Features)	0.844	0.840	0.713	0.605	1.010	0.303	0.551
Siamese+LR (OHE Features)	0.807	0.757	0.752	0.690	1.281	0.285	0.508
Siamese+LR (Target Encoded Features)	0.727	0.613	0.739	0.675	1.315	0.292	0.474
Siamese+LR (Low Data 10%)	0.829	0.817	0.766	0.710	0.994	0.288	0.500
Gaussian Noise Augmentation	0.746	0.633	0.765	0.720	1.374	0.279	0.486
Feature Masking Augmentation	0.751	0.648	0.751	0.695	1.378	0.288	0.477
Ablate Loan Features	0.768	0.667	0.658	0.490	1.221	0.326	0.483
Platt Calibration	0.714	0.601	0.763	0.715	1.426	0.283	0.458
Platt Calibration (alternate)	0.714	0.601	0.755	0.715	0.583	0.198	0.458
Isotonic Calibration	0.802	0.736	0.769	0.720	1.393	0.286	0.530
Isotonic Calibration (alternate)	0.832	0.773	0.759	0.765	0.645	0.162	0.530
Siamese+KNN (k=10) (Scaled Features)	0.744	0.655	0.753	0.695	8.425	0.273	0.467

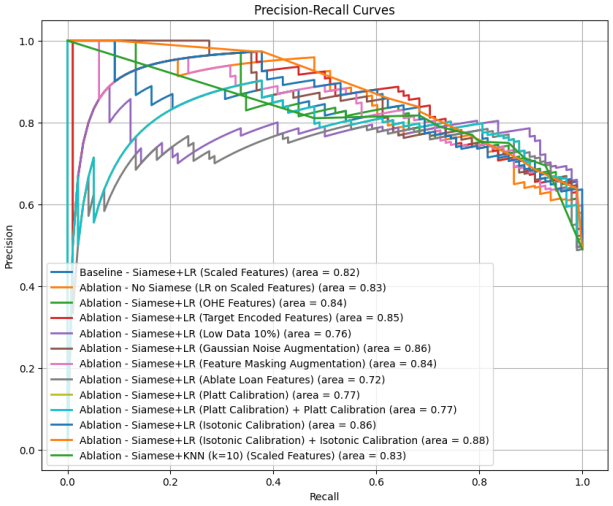

Fig. 15: Precision-Recall curves for ablation study models.

C. multi-modal features

Second, further exploration of multi-modal features—including textual loan descriptions, borrower social data, or network-based relationships—may complement existing numerical and categorical attributes, allowing models to capture richer, latent indicators of creditworthiness.

However, the method is sensitive to the construction of positive/negative pairs and to the choice of margin m, which may affect stability.

D. summary

In summary, future research should focus on improving calibration, leveraging semi- or self-supervised learning, incorporating temporal dynamics, and exploiting multi-modal data. Such approaches may enhance both overall discriminative performance and minority-class prediction, ultimately leading to more accurate, robust, and inclusive credit scoring systems in microfinance environments. This study demonstrates the effectiveness of representation learning, and in particular Siamese networks, for credit scoring in low-data and imbalanced microfinance settings. By integrating self-supervised pretraining with contrastive learning, our framework achieves competitive discriminative and calibration performance compared to strong baselines, while improving recall on minority borrower classes. These results suggest that few-shot and self-supervised methods can mitigate data scarcity challenges faced by MFIs, supporting more inclusive and reliable credit risk assessment.

Beyond predictive accuracy, the approach provides a flexible foundation for extensions such as fairness-aware modeling, domain adaptation across institutions, and meta-learning for rapidly adapting to new borrower populations. Future research will explore these directions, as well as the integration of alternative data sources, to further enhance the robustness and applicability of machine learning for financial inclusion.

REFERENCES

[1] World Bank, "Ending poverty for half the world could take more than a century," *World Bank Press Release*, 2024. [Online]. Available: https://www.worldbank.org/en/news/press-release/2024/10/15/ending-poverty-for-half-the-world-could-take-more-than-a-century. [Accessed: Sep. 10, 2025].

[2] Our World in Data, "$3 a day: A new poverty line has shifted the World Bank's data on global poverty," 2025. [Online]. Available: https://ourworldindata.org/new-international-poverty-line-3-dollars-per-day. [Accessed: Sep. 10, 2025].

[3] IPCC, "Chapter 8: Poverty, livelihoods and sustainable development," 2025. [Online]. Available: https://www.ipcc.ch/report/ar6/wg2/chapter/chapter-8/. [Accessed: Sep. 10, 2025].

[4] World Bank, "Financial inclusion overview," 2025. [Online]. Available: https://www.worldbank.org/en/topic/financialinclusion/overview. [Accessed: Sep. 10, 2025].

[5] M. Schreiner, "Credit scoring for microfinance: Can it work?," *BYU ScholarsArchive*, 2000. [Online]. Available: https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=1098&context=esr.

[6] MyBucks Malawi, "My Bucks Malawi Limited – PowerPoint presentation (AFSIC)," 2019. [Online]. Available: https://www.afsic.net/wp-content/uploads/2019/05/My-Bucks-Malawi.pdf.

[7] D. Björkegren and D. Grissen, "Behavior revealed in mobile phone usage predicts credit repayment," *arXiv preprint arXiv:1712.05840*, 2018. [Online]. Available: https://arxiv.org/pdf/1712.05840.pdf.

[8] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," *ICML Deep Learning Workshop*, 2015. [Online]. Available: http://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf.

[9] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2006.

[10] V. Bumacov and MyBucks, "Digital approaches in microfinance: The case of MyBucks Malawi Limited," *AFSIC Investing in Africa Conference*, 2017. [Online]. Available: https://www.afsic.net/wp-content/uploads/2019/05/My-Bucks-Malawi.pdf.

[11] T. Zhang et al., "Inclusive decision making via contrastive learning and domain adaptation," *SSRN*, 2024. [Online]. Available: https://papers.ssrn.com/sol3/Delivery.cfm/4496106.pdf?abstractid=4496106&mirid=1.

[12] L. Kovács et al., "Enhancing credit scoring with alternative data and machine learning for financial inclusion," *SEEJPH*, 2024. [Online]. Available: https://www.seejph.com/index.php/seejph/article/download/3584/2381/5422.

[13] World Bank, "Using psychometrics to overcome collateral constraints in Ethiopia," 2021. [Online]. Available: https://openknowledge.worldbank.org/bitstreams/1287b962-0498-5ee2-b8f1-b350b267cf12/download.

[14] D. Rahman et al., "Classification of customer credit risk levels using the random forest method," *JOCSAIC*, 2021. [Online]. Available: https://journals.raskhamedia.or.id/index.php/jocsaic/article/download/20/14.

[15] IMFEA, "Microfinance beyond group lending," 2023. [Online]. Available: https://imfea.or.id/wp-content/uploads/2023/07/Microfinance_Beyond_Group_Lending.pdf.

[16] T. Zhang et al., "Inclusive FinTech lending via contrastive learning and domain adaptation," *arXiv preprint arXiv:2305.05827*, 2023. [Online]. Available: https://arxiv.org/abs/2305.05827.

[17] LenddoEFL, "Case study: Universal credit scoring for female-led SMEs in Ethiopia," 2021. [Online]. Available: https://lenddoefl.com/news/2021/5/17/case-study-universal-credit-scoring-for-female-led-smes-in-ethiopia.

[18] M. Schreiner, "Benefits and pitfalls of statistical credit scoring for microfinance," *Microfinance.com White Paper*, 2002. [Online]. Available: https://www.microfinance.com/English/Papers/Scoring_Benefits_Pitfalls.pdf.

[19] Agbana, J. , Bukoye, J. and Arinze-Emefo, I. (2023) Impact of Credit Risk Management on the Financial Performance of Microfinance Institutions in Nigeria: A Qualitative Review. Open Journal of Business and Management, 11, 2051-2066. doi: 10.4236/ojbm.2023.115113.

[20] "Innovative Machine Learning Applications in Microfinance," *International Interdisciplinary Business Economics Advancement Journal*, vol. 5, no. 11, pp. 6–20, Nov. 2024. doi: 10.55640/business/volume05issue11-02. [Online]. Available: https://www.iibajournal.org/index.php/iibeaj/article/view/50

[21] S. Khavul, "Microfinance: Creating Opportunities for the Poor?," *Academy of Management Perspectives*, vol. 24, pp. 58–72, Aug. 2010. doi: 10.5465/AMP.2010.52842951.

[22] K. Bakshi, "Machine learning for microfinance institutions—A review," *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, no. 11, 2021.

[23] "AI-Based Credit Scoring Models in Microfinance: Improving Loan Accessibility, Risk Assessment, and Financial Inclusion," *The Critical Review of Social Sciences Studies*, vol. 3, no. 1, pp. 2997–3033, 2025. doi: 10.59075/15hhfs58. [Online]. Available: https://doi.org/10.59075/15hhfs58

[24] B. Dashnyam, G.-O. Uvgunkhuu, and B. Sosorbaram, "The Machine Learning Methods for Micro-Credit Scoring: The Case of Micro-Financing in Mongolia," *Eduvest - Journal of Universal Studies*, vol. 4, no. 5, pp. 4489–4503, 2024. doi: 10.59188/eduvest.v4i5.1137. [Online]. Available: https://doi.org/10.59188/eduvest.v4i5.1137

[25] L. Kovács *et al.*, "Enhancing Credit Scoring with Alternative Data and Machine Learning for Financial Inclusion," 2024. [Online]. Available: https://www.seejph.com/index.php/seejph/article/download/3584/2381/5422

[26] World Bank, "Using Psychometrics to Overcome Collateral Constraints in Ethiopia," 2021. [Online]. Available: https://openknowledge.worldbank.org/bitstreams/1287b962-0498-5ee2-b8f1-b350b267cf12/download

[27] Y. Li, C. L. P. Chen, and T. Zhang, "A survey on siamese network: Methodologies, applications, and opportunities," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 994–1014, 2022.

[28] L. Chioda, P. Gertler, S. Higgins, and P. C. Medina, "Fintech lending to borrowers with no credit history," National Bureau of Economic Research, Working Paper No. w33208, 2024.

[29] L. Dalla Pellegrina, D. Diriker, P. Landoni, D. Moro, and M. Wijesiri, "Financial and social sustainability in the European microfinance sector," *Small Business Economics*, vol. 63, no. 3, pp. 1249–1292, 2024. doi: 10.1007/s11187-023-00850-7. [Online]. Available: https://doi.org/10.1007/s11187-023-00850-7

[30] Y. A. Yayehyirad, "Determinants of Financial and Operational Sustainability of Selected Micro Finance Institutions in Ethiopia," *International Journal for Multidisciplinary Research (IJFMR)*, vol. 5, no. 4, Jul.–Aug. 2023. [Online]. Available: https://www.ijfmr.com

[31] I. Khanchel, N. Lassoued, and C. Khiari, "Untangling the skein: The impact of FinTech on social and financial performance in microfinance institutions," *Regional Science Policy & Practice*, vol. 17, no. 8, 100208, 2025. doi: 10.1016/j.rspp.2025.100208. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1757780225000381