Journal of Al-Farabi for Engineering Sciences (JFES)
Vol. 5, No. 1, 2026, **pp**. 196–209
ISSN: 2617-5517 | **DOI**: https://doi.org/10.59746/z76aad94

Leakage-Safe Borrower Segmentation Using Deep Clustering and Machine Learning in Large-Scale Lending Data

M. E. Alqaysi[1,*], Ahmed Subhi Abdalkafor[1], Murtadha M. Hamad[1]
[1]Department of Computer Science, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq
Email: mus22c1013@uoanbar.edu.iq; ahmed.abdalkafor@uoanbar.edu.iq; dr.mortadha61@uoanbar.edu.iq
Received 11/02/2026, Revised 18/05/2026, Accepted 23/06/2026

Abstract: Large-scale lending data often combine pre-origination borrower information with post-origination servicing and outcome fields, which can create data leakage and misleading model performance. This study proposes a leakage-safe borrower-segmentation framework for LendingClub data. The framework constructs an ex-ante analytical matrix, compares direct K-Means/Gaussian Mixture Model (GMM) baselines with deep latent-space clustering using Deep Embedded Clustering (DEC), Improved Deep Embedded Clustering (IDEC), and Deep Clustering Network (DCN), and evaluates supervised operationalization using XGBoost, Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) . The final matrix contained 2,168,488 observations and 89 features. Conventional clustering showed weak separation, with the best Silhouette (SIL) value of 0.120454, whereas DEC achieved SIL = 0.851837, Davies-Bouldin index (DBI) = 0.179926, and Calinski-Harabasz index (CH) = 1.349271 × 10^6. XGBoost achieved the strongest recoverability performance for the DEC-derived labels. The results support leakage-safe preparation followed by latent deep clustering, while supervised learning is interpreted as operational recoverability of cluster-derived labels rather than direct default prediction.
Keywords: Autoencoder; Borrower Segmentation; Deep Clustering; Ex-Ante Credit Risk; Machine Learning; Supervised Operationalization.

1. Introduction

Credit-risk analytics increasingly depends on machine-learning models capable of processing large, heterogeneous, and high-dimensional lending records [1]. In lending environments, borrower data may include loan terms, affordability indicators, credit-history variables, categorical descriptors, temporal attributes, and underwriting-related information [2]. Although such data richness can improve analysis, it also introduces sparsity, correlation, imbalance, and weakly informative variables that may distort learning if the dataset is not prepared within a clear methodological boundary [3], [4]. In ex-ante credit-risk analysis, temporal validity is essential because underwriting-oriented models must rely only on information available at or before loan origination [5]. Including post-origination variables, such as repayment behavior, servicing updates, recoveries, or outcome-dependent fields, can lead to deceptively strong results while violating the real lending-decision context [6]. Therefore, leakage-safe data preparation is a necessary foundation for credible borrower analysis.

A further challenge is that large lending datasets are not only predictive resources but also contain latent borrower structures that may not be visible through direct supervised classification. Traditional credit-risk modeling often emphasizes default prediction [7], [8], whereas segmentation-oriented analysis can reveal borrower groups with similar financial, behavioral, and application-level characteristics. Such segmentation can support portfolio understanding, borrower profiling, and operational grouping, provided that the resulting clusters are generated from leakage-safe information and interpreted without overstating their relationship to observed default outcomes [9]. Borrower segmentation is also challenging because conventional methods such as K-Means and Gaussian Mixture Model (GMM) may struggle with sparse, correlated, and mixed-type credit data [10], [11]. Latent representation learning offers an alternative by transforming the prepared borrower matrix into a compact nonlinear space before clustering [12], [13]. Despite recent progress in credit-risk machine learning and representation learning, few studies connect leakage-safe ex-ante construction, direct-versus-latent clustering comparison, and supervised recoverability of cluster-derived borrower segments within one compact workflow [14]. This gap is important

because a clustering solution may appear strong internally but still require operational testing to determine whether the learned segmentation can be reproduced at the record level by supervised models [15]. Therefore, evaluating supervised recoverability provides an additional practical check while avoiding the incorrect interpretation of cluster-derived labels as observed default-risk labels [16].

This paper compares direct conventional clustering with deep latent-space clustering using Deep Embedded Clustering (DEC), Improved Deep Embedded Clustering (IDEC), and Deep Clustering Network (DCN) [17], [18], then evaluates whether XGBoost, Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) can learn the resulting cluster-derived target classes [19], [20], [21]. The novelty is not claimed as a new clustering algorithm; rather, it lies in the integrated and leakage-aware workflow that links ex-ante data construction, direct-versus-latent clustering comparison, and supervised recoverability of cluster-derived labels. The contributions are threefold:

- Constructing a leakage-safe ex-ante analytical dataset for borrower segmentation;
- Comparing direct feature-space clustering with latent-space deep clustering as practical segmentation pipelines; and
- Evaluating supervised operationalization of DEC-derived cluster labels without claiming direct default prediction.

The remainder of the paper presents related work, methodology, results, limitations, and conclusions.

2. Related Work

2.1 Ex-Ante Credit-Risk Machine Learning

Credit-risk prediction is commonly addressed as a supervised learning problem in which borrower and loan characteristics are used to estimate repayment risk or classify creditworthiness [22], [23]. Recent studies show that boosted models, ensemble learners, and neural approaches can outperform simpler linear methods in complex credit environments because they capture nonlinear relationships and heterogeneous borrower profiles [24]. However, strong predictive performance alone does not guarantee methodological validity, especially when feature eligibility, leakage control, and borrower-structure discovery are not explicitly addressed [3], [6]. This motivates treating data preparation and segmentation as core analytical stages rather than routine preprocessing steps.

2.2 Data Leakage and High-Dimensional Credit Data

Data leakage is a critical issue in lending analytics because loan databases may contain both pre-origination variables and post-origination fields such as payment behavior, servicing updates, recoveries, and outcome-dependent indicators [5]. For ex-ante analysis, these variables must be excluded to preserve the underwriting-time decision context [25]. High-dimensional credit data also include numeric, categorical, temporal, sparse, and correlated variables that can distort feature-space geometry if they are not properly encoded, normalized, and filtered [10]. Therefore, leakage-safe construction and dimensional control are necessary before reliable clustering or supervised learning can be performed.

2.3 Clustering for Borrower Segmentation

Clustering can reveal borrower substructures that may not be captured by direct supervised prediction, since it identifies groups based on similarity rather than observed repayment labels [26]. However, conventional clustering may perform weakly in high-dimensional mixed-type credit data, where distances are affected by scale variation, redundancy, and correlated noise [27]. Representation learning provides an alternative by transforming borrower records into compact latent spaces that preserve dominant structure while reducing irrelevant variation [13]. Deep clustering extends this idea by combining learned representations with cluster refinement [18]. Few studies connect leakage-safe ex-ante construction, direct-versus-latent clustering comparison, and

supervised recoverability of the resulting segments within one workflow; this gap motivates the present study.

Table 1 provides a compact state-of-the-art positioning of the proposed workflow against closely related study directions while keeping the manuscript within the journal page limit.

Table 1. Compact positioning against related state-of-the-art study directions

Study direction	Compact positioning of this study
Credit-risk ML	Focuses on leakage-safe borrower segmentation, not only default/risk prediction.
Leakage-aware analytics	Links leakage control with clustering evaluation and supervised recoverability.
Latent clustering	Compares K-Means/GMM with DEC/IDEC/DCN and tests operational learnability.

3. Methodology

Figure 1 summarizes the proposed methodology, which is organized into three stages: leakage-safe ex-ante data preparation, clustering and algorithm comparison, and supervised operationalization with evaluation. Algorithm 1 presents the corresponding stepwise procedure.

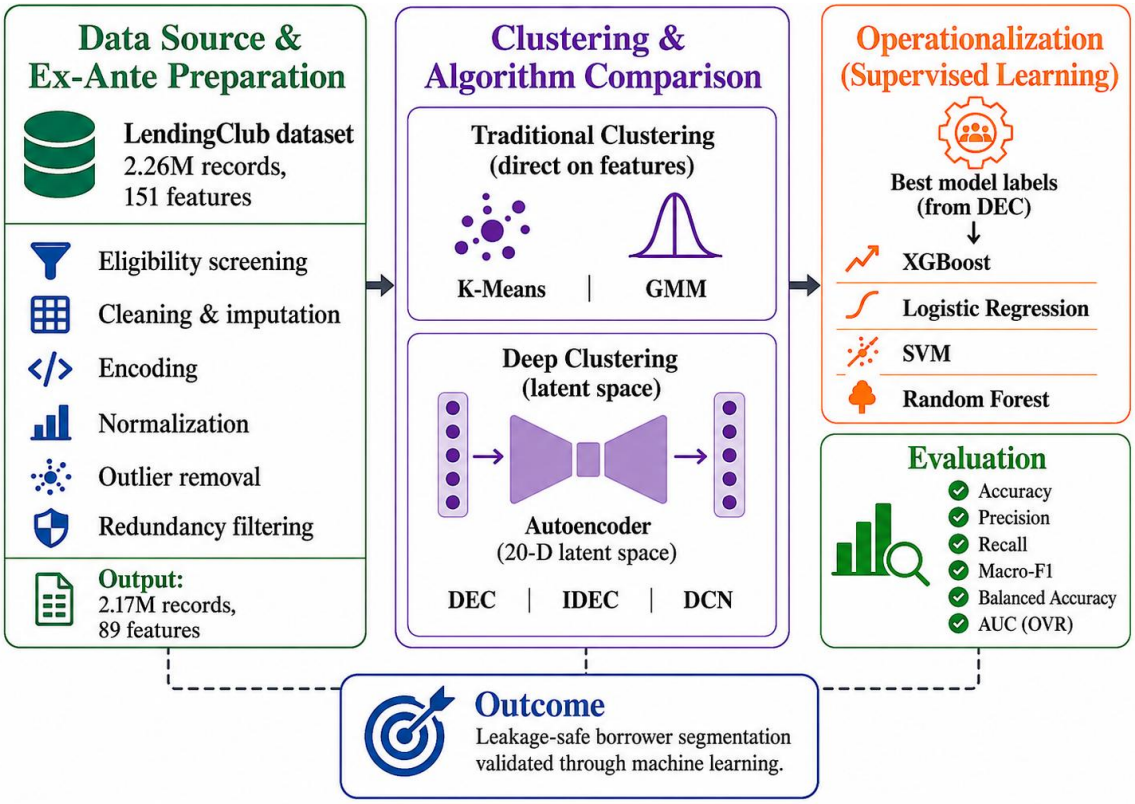

Figure 1. Proposed methodology workflow for leakage-safe borrower segmentation

Algorithm 1. Proposed leakage-safe borrower-segmentation framework

Input: Raw LendingClub loan records D_raw.
Output: Cluster-derived segmentation labels C and supervised recoverability metrics M.
Begin
1: Load D_raw and define the original borrower-feature matrix.
2: Retain only variables observable at or before loan origination.
3: Remove post-origination leakage-prone variables, identifiers, free-text fields, and non-informative administrative attributes.
4: Harmonize missing values and encode binary, ordinal, nominal, and temporal variables into a numeric analytical matrix.
5: Apply z-score normalization, PCA-based stabilization, Isolation Forest screening, and redundancy filtering to obtain X_exante.
6: Evaluate direct K-Means and GMM baselines on the prepared feature space using SIL, DBI, and CH.
7: Train an autoencoder on X_exante and extract the selected 20-dimensional latent representation Z.

8: Apply DEC, IDEC, and DCN to Z and evaluate their clustering quality using the same internal criteria.
9: Select the strongest clustering solution and assign cluster-derived labels C.
10: Train XGBoost, LR, SVM, and RF using C as the supervised target.
11: Report Accuracy, Precision, Recall, Macro-F1, Balanced Accuracy, and multiclass AUC as M.
12: Interpret M as operational recoverability of borrower segmentation, not as direct default-prediction validity.
End

3.1 Ex-Ante Data Construction and Analytical Preparation

The first stage constructs a valid ex-ante analytical dataset from the original LendingClub records, which contain 2,260,701 loan records and 151 attributes covering borrower, financial, credit-history, application, and servicing-related information. This stage is not merely a cleaning step; it establishes pipeline validity by ensuring that clustering and supervised learning use only information observable at loan origination.

The first operation is temporal eligibility control. Each variable is examined according to its time of availability, and only attributes observable at or before loan origination are retained. Variables related to repayment dynamics, servicing updates, recoveries, or post-origination outcomes are excluded before preprocessing, representation learning, clustering, and supervised modeling. This separation prevents leakage and preserves the underwriting-time decision context.

After eligibility control, retained variables are cleaned and transformed into a unified numeric matrix. Missing-value indicators, empty strings, and placeholder codes are standardized, while identifiers, free-text fields, and non-informative variables are removed. Numeric variables are imputed using robust strategies, binary variables by mode, ordinal variables through order-preserving encoding, nominal variables through one-hot encoding, and eligible dates through numeric temporal components.

The resulting matrix is standardized using z-score normalization to reduce scale dominance. PCA is used only as an intermediate stabilization space to reduce redundancy and noise before anomaly detection, not as the final representation. Isolation Forest is then applied to remove atypical observations that could distort cluster geometry, followed by low-variance and high-correlation filtering to reduce residual redundancy. Table 2 summarizes these steps and their analytical purpose. The output is a leakage-controlled ex-ante dataset of 2,168,488 observations and 89 features.

Table 2. Dataset analytical preparation summary

Component	Input / examples	Action	Output
Raw dataset	2,260,701 records; 151 attributes	Define empirical source	Original LendingClub matrix
Leakage control	loan status, recoveries, collections, last-payment and outcome-dependent fields	Remove post-origination variables	Ex-ante feature space
Non-informative fields	Identifiers, free text, administrative fields	Remove false-similarity sources	Cleaner matrix
Missingness & encoding	Numeric, binary, ordinal, nominal, temporal variables	Imputation and semantic encoding	Unified numeric matrix
Scaling & screening	Mixed scales, redundancy, atypical observations	Z-score, PCA stabilization, Isolation Forest, redundancy filtering	Stable borrower space
Final matrix	Prepared analytical data	Retain leakage-safe features	2,168,488 records; 89 features

3.2 Clustering and Algorithm Comparison

The second stage evaluates whether the prepared borrower space can be partitioned into coherent groups. The study compares two legitimate but different segmentation pipelines. The first pipeline applies conventional K-Means and GMM directly to the prepared ex-ante feature space, including raw and z-score standardized variants. These models do not use the autoencoder, and they are retained as transparent direct-clustering baselines.

The second pipeline uses an autoencoder to learn a compact 20-dimensional latent representation of the ex-ante borrower matrix, followed by deep clustering in the learned space.

The autoencoder is treated as a nonlinear representation-learning module rather than a supervised classifier; its adequacy is assessed through reconstruction behavior, validation-loss stability, reconstruction-error distribution, and downstream clustering quality. Candidate latent dimensions of 10, 15, and 20 were examined, and the 20-dimensional representation was retained because it provided the best available balance between compression and structural preservation. The deep models include DEC, IDEC, and DCN. For the conventional K-Means baseline, the number of clusters was guided using the elbow method. Based on the visible bend in the inertia curve, K = 3 was retained as a controlled and interpretable experimental setting that supports comparable evaluation across the candidate models and reflects a practical three-segment portfolio view. In this study, K = 3 is therefore treated as an elbow-supported working choice rather than as a universally optimal value.

All clustering models are evaluated using internal validation metrics. The Silhouette (SIL) score measures cohesion and separation, with larger values indicating better clustering. The Davies-Bouldin index (DBI) measures average similarity between clusters, with smaller values indicating more compact and better separated partitions. The Calinski-Harabasz index (CH) measures between-cluster dispersion relative to within-cluster dispersion, with larger values indicating stronger separation.

3.3 Supervised Machine-Learning Operationalization

In the third stage, the DEC-derived cluster identifier is used as the supervised target to evaluate whether the selected segmentation can be reproduced at the record level. The models therefore do not predict observed loan default; they learn cluster-derived borrower segmentation labels. This stage measures operational recoverability rather than default-prediction validity.

The evaluated algorithms are XGBoost, LR, SVM, and Random Forest (RF), representing boosted nonlinear learning, linear probabilistic classification, margin-based separation, and bagged tree ensembles.

The supervised models are evaluated using Accuracy, Precision, Recall, Macro-F1, Balanced Accuracy, and multiclass AUC under a one-vs-rest macro-averaging interpretation. These metrics assess overall correctness, class-level reliability, class recovery, balanced classification quality, and probability-ranking ability.

For reproducibility, Table 3 summarizes the experimental settings that are available in the compact journal version. Where implementation-level details were not preserved in the current reporting package, the manuscript avoids unsupported claims and states the limitation explicitly rather than inventing hyperparameter values.

Table 3. Compact experimental settings

Item	Compact reporting
Data	2,168,488 records; 89 ex-ante features
Baselines	K-Means/GMM on fixed 200,000-record sample
Latent space	Candidates: 10, 15, 20; selected: 20-D
Clustering	K = 3; metrics: SIL, DBI, CH
Supervised	Target: DEC labels; XGBoost, LR, SVM, RF
Limits	Full hyperparameters, seeds, profiles, and SHAP require extended reporting

4. Results and Discussion

4.1 Data Preparation Outcomes

The preprocessing stage produced a complete leakage-controlled matrix with no residual missing values after cleaning and imputation. This is important because incomplete records can distort normalization, representation learning, and cluster formation. The figures in this subsection evaluate completeness, semantic encoding, and outlier handling as the empirical foundation for the clustering comparison.

 ISSN: 2617-5517 | DOI: https://doi.org/10.59746/z76aad94

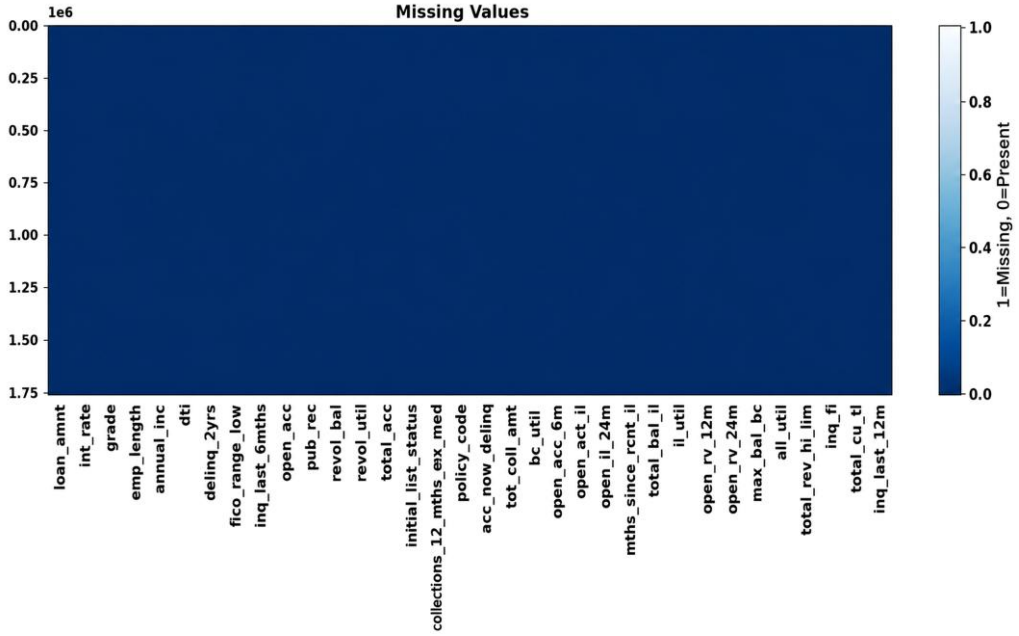

Figure 2. Missing-value status after cleaning and imputation

As shown in Figure 2, the prepared ex-ante matrix contains no residual missingness across the retained features. The uniform dark pattern indicates that missing-value handling was completed before clustering and supervised learning, reducing the risk that the results are driven by incomplete records or inconsistent imputation artifacts.

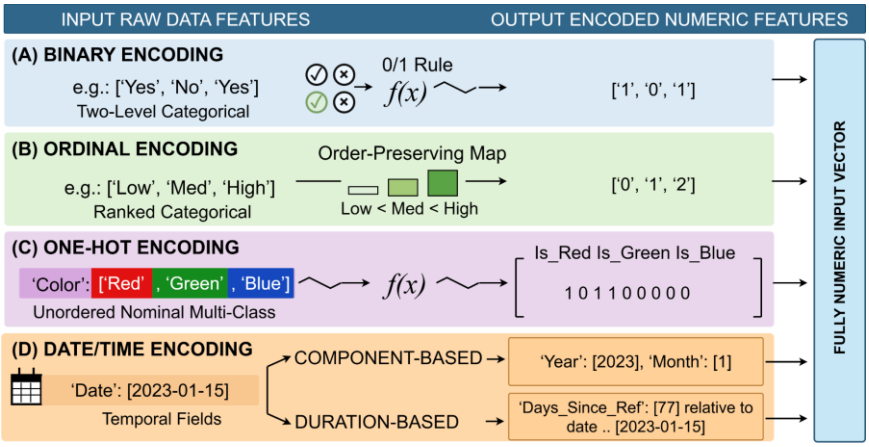

Figure 3. Encoding framework for heterogeneous ex-ante variables

As shown in Figure 3, the semantic encoding framework converts heterogeneous raw variables into a fully numeric input vector. Binary variables are mapped directly, ordinal variables preserve their ranking, nominal variables are one-hot encoded, and temporal variables are transformed into numerical components, reducing the risk of artificial borrower similarities caused by inappropriate coding.

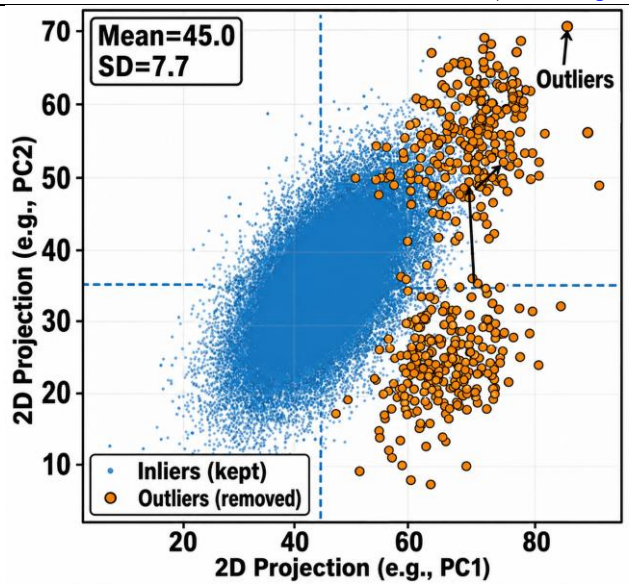

Figure 4. Isolation Forest outlier-screening results

As shown in Figure 4, the Isolation Forest screening is targeted rather than destructive: removed observations are concentrated in low-density outer regions, while the central borrower mass is preserved. This improves geometric coherence for clustering without erasing the dominant borrower population.

4.2 Conventional Clustering Baselines

Before evaluating deep clustering, direct conventional baselines were examined to establish how well K-Means and GMM could segment the prepared feature space without autoencoder-based representation learning. Table 4 reports the conventional configurations applied to raw and z-score standardized versions of the ex-ante matrix. This section is placed before the deep-clustering diagnostics to clarify the baseline level of structural separation achieved by traditional clustering alone.

Due to computational constraints, K-Means and GMM were evaluated on a fixed 200,000-record sample, whereas the deep clustering results represent the final latent-space solution. Therefore, the comparison is interpreted as a practical segmentation-pipeline comparison rather than a strictly identical-sample algorithmic benchmark. This framing is intentionally conservative: the results support the usefulness of the latent deep-clustering route in this workflow, but they do not claim that DEC is universally superior under every possible identical-input experimental design.

Table 4. Conventional K-Means and GMM results

Model	SIL	DBI	CH	Cluster sizes
KM-raw	0.073	2.673	20299.279	[77,543; 6,835; 115,622]
GMM-raw	0.120	5.612	14097.439	[64,724; 8,665; 126,611]
KM-z	0.075	2.742	18678.917	[76,743; 6,740; 116,517]
GMM-z	0.120	3.187	9534.375	[26,736; 16,743; 156,521]

The conventional results in Table 4 indicate limited separation. The best SIL score among these models is 0.120454 for GMM-z, suggesting weak cohesion and separation. K-Means performs poorly in both raw and standardized settings, indicating that scale normalization alone is insufficient for discovering a stable borrower partition. GMM improves SIL slightly, but the DBI values remain high, especially in the raw feature space, which indicates overlapping or poorly separated mixture components. The uneven cluster-size distributions further suggest that direct feature-space clustering is useful as a transparent baseline but insufficient as the main segmentation route.

4.3 Deep-Clustering Diagnostics

After establishing the conventional baseline, the next step examined whether autoencoder-based representation learning produced a latent space suitable for deep clustering. The selected 20-dimensional representation was retained because it balanced compression, reconstruction fidelity, and visible structural organization. The following diagnostics evaluate convergence, reconstruction quality, neighborhood structure, and DEC optimization behavior. As shown in Figure 5, both training and validation losses decline steadily, indicating that the autoencoder learned a usable representation rather than merely memorizing the training data. This supports the use of the 20-dimensional latent space as the basis for deep clustering.

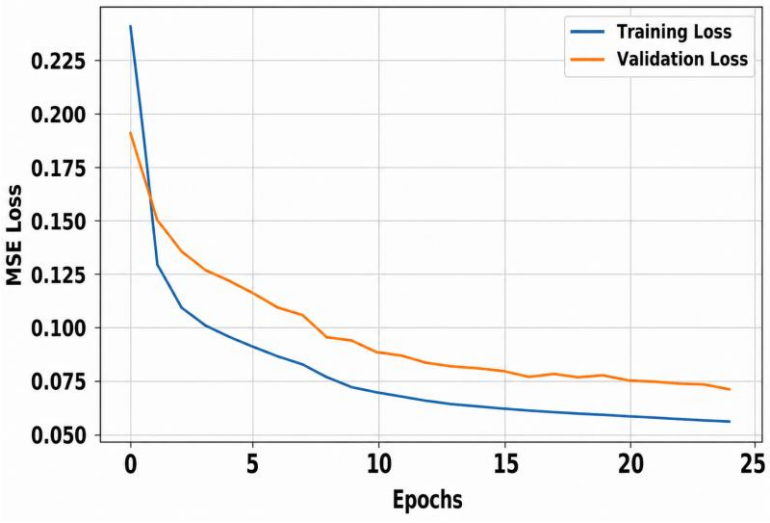

Figure 5. Autoencoder training and validation loss

As shown in Figure 6, most observations have low reconstruction error, while only a small portion exhibits higher error. This pattern suggests that the 20-dimensional latent representation retains the dominant borrower structure, with the high-error tail reflecting atypical or complex borrower profiles rather than a general representational failure.

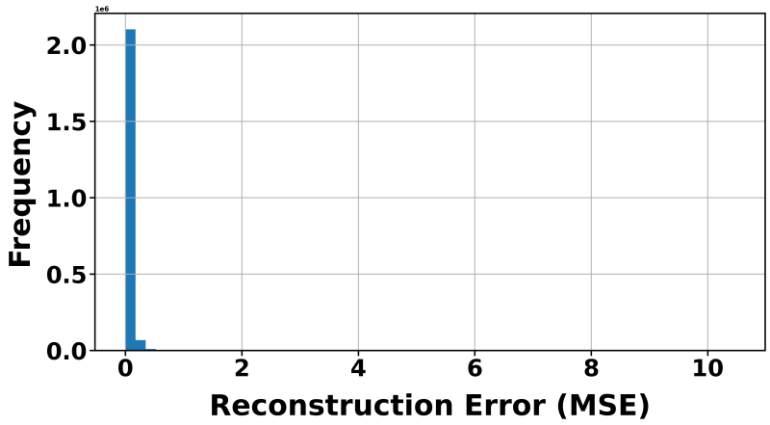

Figure 6. Reconstruction-error distribution of the autoencoder

As shown in Figure 7, the elbow curve exhibits a pronounced reduction in inertia from K = 2 to K = 3, followed by a more gradual decline for larger values of K. This pattern indicates that the three-cluster solution provides a reasonable balance between within-cluster compactness and model simplicity. Accordingly, K = 3 was retained as the experimental setting used for the clustering comparisons reported in this study.

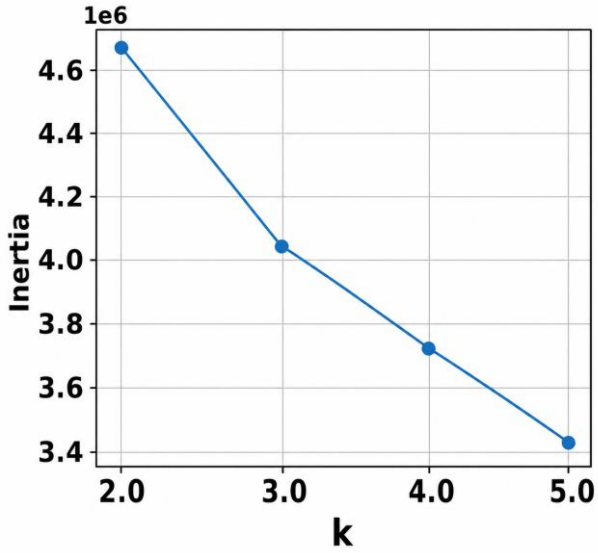

Figure 7. Elbow method for selecting the number of clusters

As shown in Figure 8, the DEC loss drops sharply at the beginning and then moves toward a more stable region. This pattern indicates rapid refinement from an informative initial partition, while later non-monotonic behavior is expected because soft assignments and target distributions are updated iteratively.

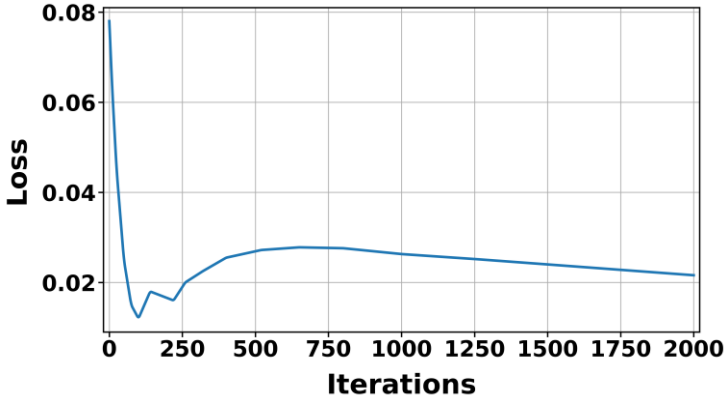

Figure 8. DEC training loss curve across clustering iterations

4.4 Deep Clustering Results and Comparison with Traditional Clustering

The deep clustering results show a stronger partitioning structure than the conventional baselines. As reported in Table 5, DEC achieved the best values across all three internal validation metrics: SIL = 0.851837, DBI = 0.179926, and CH = 1.349271 × 10^6. DCN ranked second, while IDEC ranked third. Because the baselines and deep models use different segmentation pipelines, the results should be interpreted as practical pipeline evidence rather than a strictly identical-input benchmark.

Table 5. Deep clustering performance

Model	SIL	DBI	CH
DEC	0.851837	0.179926	1.349271e6
IDEC	0.451576	0.778803	1.922603e5
DCN	0.686198	0.383358	4.517298e5

These results suggest that DEC benefited most from the learned latent geometry. Its target-distribution refinement mechanism sharpens soft assignments after an initial clustering structure has been formed. IDEC, which preserves reconstruction regularization during clustering, may have maintained representational fidelity at the expense of stronger separation under the present data

 ISSN: 2617-5517 | DOI: https://doi.org/10.59746/z76aad94

conditions. DCN produced a stronger result than IDEC but remained below DEC, suggesting that centroid-aware representation learning was useful but less effective than DEC refinement for this borrower space. As shown in Figure 9, DEC provides the highest SIL score by a wide margin. The conventional K-Means and GMM baselines remain close to zero, indicating weak cohesion and separation, while IDEC and DCN improve over the traditional baselines but remain below DEC.

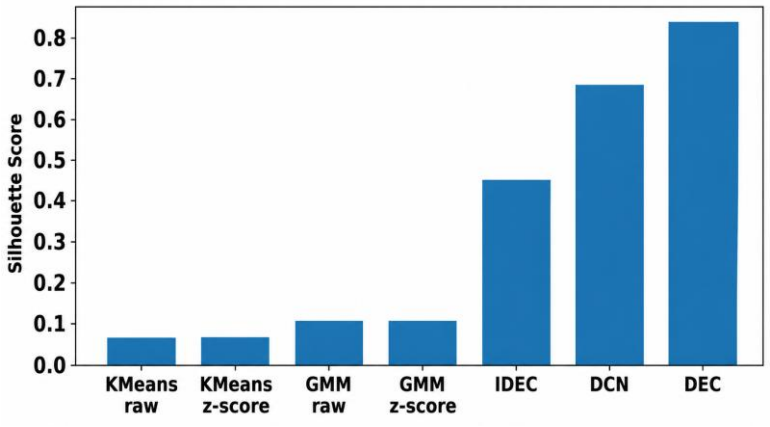

Figure 9. SIL comparison between clustering models

As shown in Figure 10, the DBI results confirm the same conclusion from the opposite direction. DEC achieves the lowest DBI, indicating better compactness and separation, whereas the conventional baselines show higher DBI values and therefore weaker internal cluster structure.

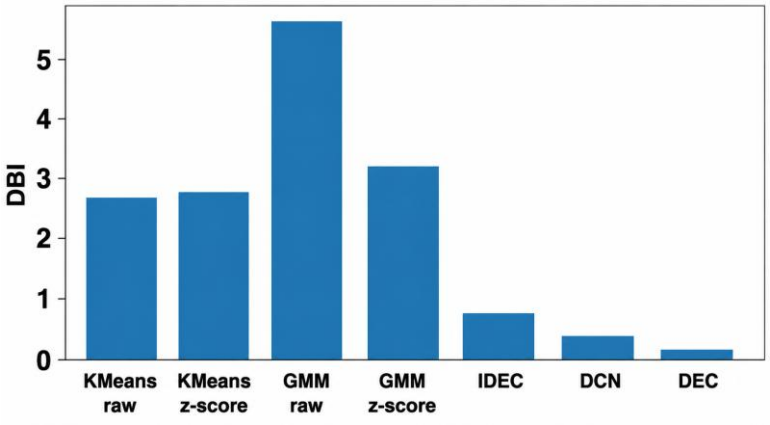

Figure 10. DBI comparison between clustering models

As shown in Figure 11, the logarithmic CH comparison highlights the magnitude of DEC improvement. Since CH reflects between-cluster dispersion relative to within-cluster dispersion, the much larger DEC value indicates a substantially stronger separation structure than both conventional and alternative deep clustering models.

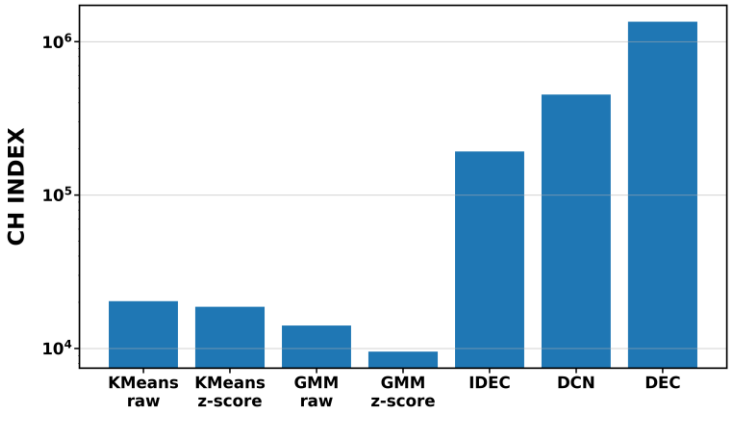

Figure 11. CH comparison between clustering models

As shown in Figure 12, the cluster-size distribution provides a plausibility check beyond metric values. The conventional settings produce highly uneven partitions, whereas DEC avoids extremely degenerate separation and yields three large borrower groups, making it more defensible for supervised operationalization.

Taken together, the figures show that the difference between traditional and deep clustering is not marginal. Using the best conventional result as the baseline, DEC increases the SIL score from 0.120454 to 0.851837, an improvement of approximately 7.07 times. DEC also reduces the best conventional DBI from 2.672802 to 0.179926, indicating far better compactness and separation. For CH, DEC reaches 1.349271 × 10^6 compared with the best conventional value of 20,299.278772, which is approximately 66.47 times larger. These results show that the deep latent-space clustering pipeline provides a substantially stronger borrower segmentation than direct traditional clustering on the prepared feature matrix.

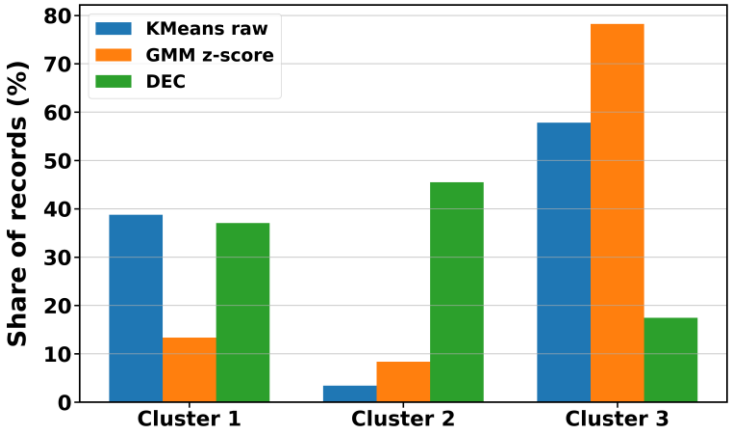

Figure 12. Cluster-size distribution

Table 6 summarizes the best conventional clustering result against DEC to provide a direct metric-by-metric comparison between the strongest traditional baseline and the selected deep clustering model. Although the internal metrics favor DEC, external credit-risk validity should be examined through post-hoc cluster profiling using ex-ante characteristics such as loan amount, interest rate, income, debt-to-income ratio (DTI), grade, purpose, and employment length. The present compact version does not assign semantic labels such as low-, medium-, or high-risk to the clusters without those profile statistics, because doing so would over-interpret an unsupervised segmentation. This analysis is therefore identified as a priority for future empirical extension.

Table 6. Best conventional clustering result versus DEC

Metric	Best conv. value	Best conv. model	DEC value	Interpretation
SIL	0.120454	GMM-z	0.851837	Stronger cluster separation
DBI	2.672802	KM-raw	0.179926	Lower DBI indicates better compactness
CH	20,299.278772	KM-raw	1.349271e6	Higher CH indicates stronger separation

4.5 Machine-Learning Results

The final empirical stage evaluates whether the selected segmentation can be learned by supervised models. This step is important because a clustering output may be internally coherent but still difficult to reproduce at the record level. In this stage, the cluster-derived class is used as the target variable for supervised operationalization. Therefore, the results should be interpreted as evidence of recoverability of the selected segmentation structure, not as direct default-prediction performance.

Table 7 reports supervised recoverability results for XGBoost, LR, SVM, and RF. XGBoost achieved the strongest performance, with Accuracy = 0.966, Precision = 0.958, Recall = 0.966, Macro-F1 = 0.962, Balanced Accuracy = 0.966, and AUC = 0.998. Because the supervised target is derived from clusters, these metrics indicate recoverability of the selected segmentation rather than

true default-prediction validity. Future extended evaluations should include a confusion matrix and class-level precision, recall, and F1 scores to verify that performance is not dominated by the largest cluster. The supervised comparison shows that XGBoost is the strongest operational model, suggesting that nonlinear boosted interactions are well aligned with the selected borrower segmentation. LR and SVM are close in overall performance, while RF shows lower classification scores but remains competitive in AUC. Thus, the segmentation is not merely a geometric output; it can be translated into a supervised target for operational use.

Table 7. Supervised machine-learning performance

Model	Accuracy	Precision	Recall	Macro-F1	Balanced Accuracy	AUC
XGBoost	0.966	0.958	0.966	0.962	0.966	0.998
LR	0.906	0.885	0.901	0.892	0.901	0.979
SVM	0.905	0.889	0.889	0.889	0.889	0.962
RF	0.890	0.880	0.879	0.879	0.879	0.974

4.6 Integrated Analytical Interpretation

Across the three stages, the findings support a coherent analytical workflow. Leakage-safe preparation ensures that clustering and supervised learning use information available at loan origination. Conventional K-Means and GMM provide transparent baselines, but DEC on the learned latent representation substantially improves internal clustering quality. In practice, the framework could support borrower profiling, portfolio segmentation, underwriting-time grouping, and risk-based customer management, provided that fairness, transparency, and responsible use of borrower data are maintained.

4.7 Limitations

The study has four main limitations. First, the supervised labels are DEC-derived clusters rather than observed default outcomes; therefore, the supervised stage measures segmentation recoverability, not default-risk prediction. Second, the conventional baselines were evaluated on a fixed sample, so the comparison should be interpreted as a practical segmentation-pipeline comparison rather than a strict identical-sample benchmark. Third, validation mainly relies on internal clustering metrics, while full post-hoc cluster profiles are not reported in this compact version. Fourth, full autoencoder hyperparameters, repeated-seed results, confusion matrices, class-level diagnostics, and SHAP explanations are not reported as completed results due to page-limit and reporting constraints. Future work should address external default-outcome validation, cluster profiling, k-sensitivity, robustness testing, detailed hyperparameter reporting, class-level evaluation, and explainability analysis.

5. Conclusion

This paper presented a compact framework for ex-ante lending analytics that integrates leakage-safe data preparation, borrower clustering, and supervised operationalization. Its contribution is a workflow that constructs a valid ex-ante feature space, compares direct clustering with deep latent-space clustering, and tests whether the selected segmentation can be reproduced by supervised models. The preparation stage produced a leakage-safe dataset of 2,168,488 observations and 89 features by excluding post-origination variables, handling missingness, encoding heterogeneous attributes, standardizing features, screening outliers, and reducing redundancy. The clustering comparison showed that the deep latent-space pipeline was stronger than direct conventional clustering. K-Means and GMM produced weak results, whereas DEC achieved SIL = 0.851837, DBI = 0.179926, and CH = 1.349271 × 10^6, supporting the value of learning a nonlinear borrower representation before cluster refinement. Practically, the framework can support borrower profiling and portfolio segmentation before supervised deployment. XGBoost provided the strongest recoverability of the selected cluster-derived labels; however, these labels are not observed default outcomes. Future work should validate clusters against external credit-risk outcomes, test stability across datasets and random seeds, perform k-sensitivity analysis, add

borrower-level cluster profiles, report class-level supervised diagnostics, and introduce SHAP or feature-importance analysis for operational interpretability.

Acknowledgements

The authors would like to thank the editorial office and reviewers for their valuable comments and constructive suggestions.

Funding Information

This research received no external funding.

Author Contributions

M. E. Alqaysi: conceptualization, methodology, data preparation, analysis, visualization, and writing - original draft. Ahmed Subhi Abdalkafor: supervision, validation, and writing - review and editing. Murtadha M. Hamad: supervision, project administration, validation, and writing - review and editing. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Ethical Approval

This study did not involve human or animal subjects, and ethical approval was not required.

Data Availability Statement

The data supporting the findings of this study are based on the public LendingClub loan dataset. (https://www.kaggle.com/datasets/wordsforthewise/lending-club)

References

[1] S. H. Goldmann, M. R. Machado, and J. R. Osterrieder, "Advancing credit risk assessment in the retail banking industry: A hybrid approach using time series and supervised learning models," *Data Knowl. Eng.*, p. 102490, 2025.

[2] T. Miljkovic and P. Wang, "A dimension reduction assisted credit scoring method for big data with categorical features," *Financial Innovation*, vol. 11, no. 1, pp. 1–30, 2025, doi: 10.1186/s40854-024-00689-1.

[3] K. Hayat and B. Magnier, "Data Leakage and Deceptive Performance: A Critical Examination of Credit Card Fraud Detection Methodologies," *Mathematics*, vol. 13, no. 16, pp. 1–28, 2025, doi: 10.3390/math13162563.

[4] S. R. Lenka, S. K. Bisoy, and R. Priyadarshini, "Multiple optimized ensemble learning for high-dimensional imbalanced credit scoring datasets," *Knowl. Inf. Syst.*, vol. 66, no. 9, pp. 5429–5457, 2024, doi: 10.1007/s10115-024-02129-z.

[5] X. Yang, H. Xue, Q. Hu, and Y. Zhang, "Design of a full-cycle intelligent risk control system for pre-loan, mid-loan, and post-loan lending: AI-driven closed-loop management of online credit security," in *Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science*, 2025, pp. 1022–1027.

[6] Q. Fan, "Risk Prediction Model of Financial Lending Big Data Leakage Based on Association Rules," in *Smart Innovation, Systems and Technologies*, T. Wang, S. Patnaik, W. C. Ho Jack, and M. L. Rocha Varela, Eds., Singapore: Springer Nature Singapore, 2023, pp. 617–629. doi: 10.1007/978-981-19-2768-3_60.

[7] J. P. Noriega, L. A. Rivera, and J. A. Herrera, "Machine learning for credit risk prediction: A systematic literature review," *Data (Basel).*, vol. 8, no. 11, p. 169, 2023.

[8] V. Chang, S. Sivakulasingam, H. Wang, S. T. Wong, M. A. Ganatra, and J. Luo, "Credit risk prediction using machine learning and Deep Learning: A study on Credit card customers. Risks, 12 (11), 174," 2024.

[9] F. Baser, O. Koc, and A. S. Selcuk-Kestel, "Credit risk evaluation using clustering based fuzzy classification method," *Expert Syst. Appl.*, vol. 223, p. 119882, 2023.

[10] Z. Wang, Q. Li, H. Zhao, and F. Nie, "Simultaneous local clustering and unsupervised feature selection via strong space constraint," *Pattern Recognit.*, vol. 142, p. 109718, 2023, doi: https://doi.org/10.1016/j.patcog.2023.109718.

[11] W. Lippitt, N. E. Carlson, J. Arbet, T. E. Fingerlin, L. A. Maier, and K. Kechris, "Limitations of clustering with PCA and correlated noise," *J. Stat. Comput. Simul.*, vol. 94, no. 10, pp. 2291–2319, 2024, doi: 10.1080/00949655.2024.2329976.

[12] T. Li, G. Kou, and Y. Peng, "A new representation learning approach for credit data analysis," *Inf. Sci. (N. Y).*, vol. 627, pp. 115–131, 2023, doi: 10.1016/j.ins.2023.01.068.

[13] R. A. Mancisidor, M. Kampffmeyer, K. Aas, and R. Jenssen, "Learning latent representations of bank customers with the variational autoencoder," *Expert Syst. Appl.*, vol. 164, p. 114020, 2021.

[14] S. Chantamunee, P. Thamrongrat, P. Thanathamathee, K. Chaisriya, and D. N. M. Nizam, "Unsupervised Deep Clustering With Hard Balanced Constraint: Application in Disciplinary-Focused Student Section Formation," *IEEE Access*, vol. 12, no. April, pp. 98239–98253, 2024, doi: 10.1109/ACCESS.2024.3423807.

[15] B. V Sánchez Vinces, E. Schubert, A. Zimek, and R. L. F. Cordeiro, "A comparative evaluation of clustering-based outlier detection," *Data Min. Knowl. Discov.*, vol. 39, no. 2, p. 13, 2025.

[16] B. Lafabregue, J. Weber, P. Gançarski, and G. Forestier, "End-to-end deep representation learning for time series clustering: a comparative study," *Data Min. Knowl. Discov.*, vol. 36, no. 1, pp. 29–81, 2022, doi: 10.1007/s10618-021-00796-y.

[17] M. Sadeghi and N. Armanfard, "Deep Multirepresentation Learning for Data Clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15675–15686, 2024, doi: 10.1109/TNNLS.2023.3289158.

[18] K. Wang, M. Li, J. Cheng, X. Zhou, and G. Li, "Research on personal credit risk evaluation based on XGBoost," *Procedia Comput. Sci.*, vol. 199, pp. 1128–1135, 2021, doi: 10.1016/j.procs.2022.01.143.

[19] H. Li, "Comparative Study of Personal Credit Default Risk Prediction Based on Different Machine Learning Models," in *ITM Web of Conferences*, EDP Sciences, 2025, p. 1016.

[20] X. Huang, "Research on Credit Risk Prediction Technology Based on Deep Learning," in *Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition*, 2024, pp. 270–274.

[21] P. Kündig and F. Sigrist, "A Spatio-Temporal Machine Learning Model for Mortgage Credit Risk: Default Probabilities and Loan Portfolios," *Eur. J. Oper. Res.*, 2025.

[22] M. Gramifar, G. Kabir, and S. A. Khan, "Credit risk assessment using bayesian networks and machine learning approaches," *Advanced Engineering Informatics*, vol. 74, p. 104638, 2026.

[23] T. Liu and D. Huang, "Research on credit risk assessment based on machine learning," in *Fifth International Conference on Mechatronics and Computer Technology Engineering (MCTE 2022)*, SPIE, 2022, pp. 1420–1425.

[24] L. Yu, L. Yu, and K. Yu, "A high-dimensionality-trait-driven learning paradigm for high dimensional credit classification," *Financial Innovation*, vol. 7, no. 1, 2021, doi: 10.1186/s40854-021-00249-x.

[25] L. Chen, L. Sun, and C. Zhong, "An integrated interpretation and clustering model based on attribute grouping," *Applied Intelligence*, vol. 55, no. 7, p. 599, 2025, doi: 10.1007/s10489-025-06262-2.

[26] J. Bosker, M. Gürtler, and M. Zöllner, "Machine learning-based variable selection for clustered credit risk modeling," *Journal of Business Economics*, vol. 95, no. 4, pp. 617–652, 2025, doi: 10.1007/s11573-024-01213-8.