A Unified Framework for Modeling Heterogeneous Financial Data via Dual-Granularity Prompting

Yu Lei [1,2], Zixuan Wang [1*], Yiqing Feng [1,2], Junru Zhang [3], Yahui Li [2]
Chu Liu [1], Tongyao Wang [1], Dongyang Li [1]

[1] DiDi International Business Group
[2] Beijing University of Posts and Telecommunications [3] Zhejiang University

Abstract

Recent industrial credit scoring models remain heavily reliant on manually tuned statistical learning methods. Despite their potential, deep learning architectures have struggled to consistently outperform traditional statistical models in industrial credit scoring, largely due to the complexity of heterogeneous financial data and the challenge of modeling evolving creditworthiness. To bridge this gap, we introduce FinLangNet, a novel framework that reformulates credit scoring as a multi-scale sequential learning problem. FinLangNet processes heterogeneous financial data through a dual-module architecture that combines tabular feature extraction with temporal sequence modeling, generating probability distributions of users' future financial behaviors across multiple time horizons. A key innovation is our dual-prompt mechanism within the sequential module, which introduces learnable prompts operating at both feature-level granularity for capturing fine-grained temporal patterns and user-level granularity for aggregating holistic risk profiles. Notably, real world deployment yielded a 6.3 pp improvement in KS, along with a 9.9% reduction in bad debt rate.

1 Introduction

Credit risk prediction is a cornerstone for financial institutions to devise effective lending policies and informed decisions evaluating the solvency of borrowers (Genovesi et al., 2023). This process is critical in minimizing loan default risks, which is essential for preserving low bad debt levels and mitigating financial losses in the multi-billion dollar credit industry (Cheng et al., 2020). Credit risk models perform binary classifications to discern good from bad customers, improving overdue risk prediction, and effectively managing bad debt while maintaining profitability (Zhao et al., 2023).

In industrial credit risk assessment, user data typically originates from multiple heterogeneous sources including credit reports, transaction histories, and product usage behaviors (Lu and Zhang, 2023). These multi-source data present significant challenges: they are inherently noisy, high-dimensional, sparse, and discrete, often with substantial missing values (Elia et al., 2023). Furthermore, user creditworthiness is not static but evolves dynamically over time, requiring models to capture both short-term behavioral changes and long-term credit profile evolution (Niazkar et al., 2024).

The risk control industry predominantly relies on XGBoost for its stability and interpretability when handling irregular multisource financial data (Li et al., 2022). The standard practice involves extensive feature engineering to create derivative features, followed by feature selection and XGBoost modeling (Song et al., 2023). While this approach achieves strong performance, it suffers from critical limitations: (1) extensive feature engineering is time-consuming and requires substantial domain expertise; (2) static models fail to capture temporal dependencies in sequential user behaviors; (3) point-in-time predictions cannot model the dynamic evolution of creditworthiness across different time horizons.

Recent advancements in time-series (Yao et al., 2022), sequential models (Yousefi and Tosarkani, 2022), and graph models (Xue et al., 2024) have shown promise in capturing temporal dynamics. However, these methods still focus on point-in-time predictions, failing to model how user creditworthiness evolves across multiple future time horizons. While XGBoost performs well for immediate risk assessment, its effectiveness deteriorates

* Corresponding Authors.

for longer-term predictions. In practical scenarios, understanding user behavior across various future time windows is crucial for comprehensive risk management. Therefore, we reformulate credit scoring as a multi-scale behavioral representation learning problem, where the model learns to characterize user profiles across different future periods.

In this work, we propose FinLangNet, a novel framework that reformulates credit scoring as a multi-scale sequential learning problem. Drawing inspiration from transformer architectures' success in capturing long-range dependencies, FinLangNet processes heterogeneous financial data through a dual-module approach: (1) a non-sequential module for extracting high-order interactions from tabular features, and (2) a sequential module with an innovative dual-prompt mechanism. This mechanism introduces feature-level prompts for capturing fine-grained temporal patterns and user-level prompts for aggregating holistic risk profiles. Through multi-scale predictions with dynamically weighted loss functions, FinLangNet generates comprehensive representations that capture both static characteristics and evolving behavioral dynamics.

Our contributions are summarized as follows:

1. We reformulate credit scoring from classification to dynamic multi-scale forecasting, enabling the model to generate future behavioral distributions that capture evolving creditworthiness patterns across different time horizons.
2. We propose FinLangNet [1], a hybrid architecture with two complementary modules: a DeepFM-based non-sequential module for tabular data and a sequential module featuring an innovative dual-prompt mechanism. The dual-prompt design operates at feature-level and user-level granularities, enabling comprehensive representation learning from heterogeneous financial data.
3. We achieve state-of-the-art results on multiple benchmarks. On the public UEA time series classification benchmark, FinLangNet outperforms existing methods by substantial margins, demonstrating its superiority as a general-purpose sequential modeling framework. The model also shows consistent improvements across various financial risk assessment metrics in real-world datasets.
4. We deployed FinLangNet in our finance platform. It achieved a 6.3 pp absolute gain in KS and 9.9% relative reduction in bad debt rate compared to the previous XGBoost-based system, demonstrating both superior risk discrimination capability and substantial financial impact in real-world industrial settings.

[1] https://github.com/didiglobal-fintech-credit-risk/FinLangNet

2 Preliminaries

Our goal is to learn a representation for predicting user credit risk by leveraging heterogeneous user data. We formulate the input as $X = (m, z)$, consisting of two modalities. The non-sequential component $m \in \mathbb{R}^M$ represents static attributes (e.g., user profiles) where M is the feature dimension. The sequential component $z = \{z_1, \ldots, z_S\}$ captures dynamic behaviors from S different sources (e.g., query records, loan logs). Each source $z_s \in \mathbb{R}^{C_s \times T_s}$ contains multivariate quantitative features with channel size C_s and sequence length T_s, which may vary across sources. To enable sequential processing, we process all temporal sources in chronological order. Visual illustrations and detailed specifications of these inputs are provided in Appendix A.

The prediction task is defined as learning a function $f_\theta : X \to [0, 1]^L$. The output is a probability distribution over L different time scales, where each entry represents the likelihood of overdue behavior within a specific future horizon. The model is trained under supervision using ground-truth labels that indicate whether a default event occurred at the corresponding time scale, guiding the model to capture risk dynamics over time.

3 Methodology

In this section, we present FinLangNet, a dual-module framework designed to capture the distinct characteristics of heterogeneous financial data. As illustrated in Figure 1, our framework consists of two complementary branches: a **Non-Sequential Module** for static user profiles and a **Sequential Module** (SRG) for temporal behaviors. The outputs from both modules are fused into a unified embedding to generate risk predictions across multiple future time scales.

3.1 Non-Sequential Module

Motivation. Non-sequential features (e.g., user profiles) contain complex static interactions that are crucial for defining baseline risk levels. Simple linear models often fail to capture the intricate com-

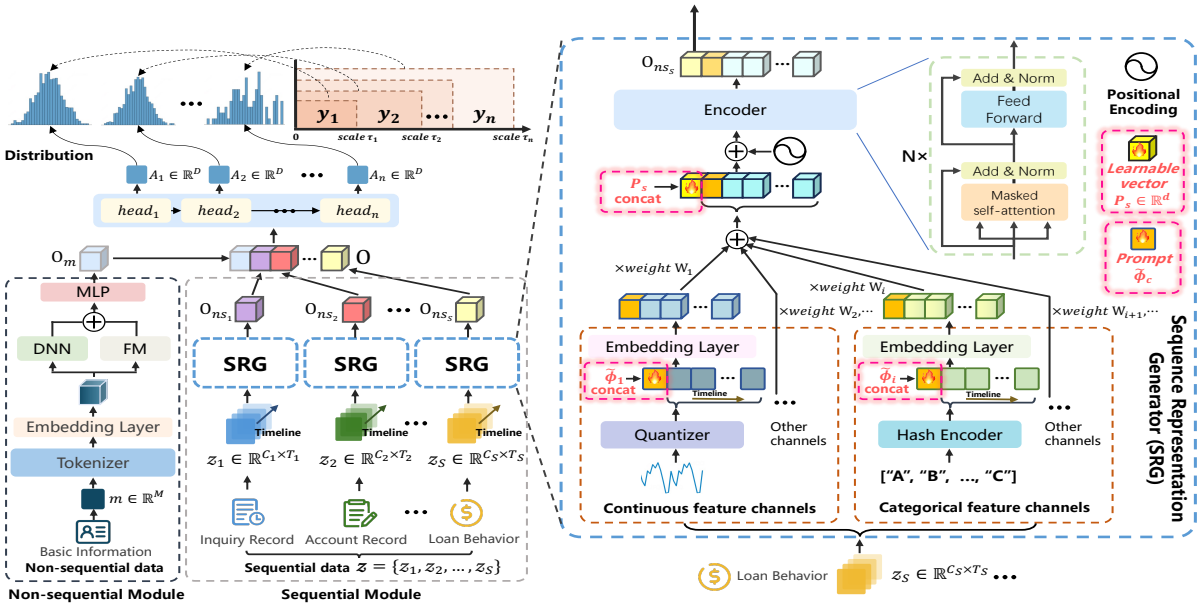

Figure 1: Overview of the FinLangNet Framework. It processes static data via a Non-Sequential Module and temporal data via a Sequence Representation Generator (SRG). The SRG utilizes a dual-prompt mechanism to capture both feature-level and user-level patterns.

binations of these attributes (e.g., the combined risk of age, occupation, and income level).

Design. We adopt DeepFM (Guo et al., 2017) to capture complex interactions within the static feature vector $m \in \mathbb{R}^M$. DeepFM combines a Factorization Machine (FM) component for low-order (second-order) feature interactions and a deep neural network (DNN) component for high-order non-linear correlations. Concretely, the FM component is defined as

$$y_{\text{FM}} = \langle w, m \rangle + \sum_{j_1=1}^{M} \sum_{j_2=j_1+1}^{M} \langle V_{j_1}, V_{j_2} \rangle m_{j_1} m_{j_2}, \tag{1}$$

where w denotes linear weights and V_j is the latent embedding for the j-th feature. In parallel, the DNN component takes the aggregated embedding signal and models higher-order interactions via an MLP:

$$y_{\text{DNN}} = \text{MLP}\left(\frac{1}{2}\left[\left(\sum_{i=1}^{M} m_i V_i\right)^2 - \sum_{i=1}^{M} (m_i^2 V_i^2)\right]\right). \tag{2}$$

We combine the two parts to obtain the static interaction embedding:

$$O_m = f_{\text{DeepFM}}(m) = \sigma\left(y_{\text{FM}} + y_{\text{DNN}}\right), \tag{3}$$

where $\sigma(\cdot)$ is a sigmoid function.

3.2 Sequential Module (SRG)

Motivation. Financial sequences differ significantly from natural language: they are multi-source, highly sparse, and contain noise. Standard RNNs often struggle with long-term dependencies in such heterogeneous data. To address this, we propose the Sequence Representation Generator (SRG), which transforms continuous financial signals into discrete tokens and captures dependencies via a novel dual-prompt mechanism.

Discretization. First, to mitigate data sparsity and robustness against noise, we discretize continuous features. For each channel c in a source z_s, a tokenizer $\mathcal{T}_c$ transforms the continuous vector into discrete tokens $t_c \in \mathbb{N}^{T_s}$.

Dual-Prompt Mechanism. Inspired by the '[CLS]' token in functional Transformers (Devlin et al., 2019), we introduce prompts at two granularities to guide representation learning:

- **Feature-level Prompt ($\tilde{\phi}_c$):** Captures channel-specific global patterns. We prepend a learnable token to each channel sequence: $t'_c = (\tilde{\phi}_c, t_{c,1}, \dots, t_{c,T_s})$. After embedding, the prompt $\tilde{\phi}_c$ aggregates information unique to that feature dimension.
- **User-level Prompt (P_s):** Captures holistic user behavior across all channels. We

aggregate channel-wise representations via weighted attention to form a unified sequence H', and prepend a global learnable vector P_s.

The augmented sequence is processed by a Transformer encoder. The state of the user-level prompt at the final layer serves as the comprehensive sequential representation $O_{\text{ns}_s} = H^{(L)}[0]$. We concatenate representations from all S sources to obtain O_{ns}.

3.3 Multi-scale Credit Risk Prediction

Motivation. Credit risk is dynamic; a user might be safe in the short term but risky in the long term. A single prediction point is insufficient for comprehensive risk management. Furthermore, financial data suffers from severe class imbalance and varying sample difficulty.

Prediction & Optimization. We fuse the static and sequential embeddings into a shared representation $O = [O_m; O_{\text{ns}}]$. To predict risks over n different horizons (e.g., 30, 60, 90 days), we employ a multi-task setup where specific heads project O to binary probabilities $\{y'_1, \ldots, y'_n\}$. To robustly train this model, we design a Dynamically Weighted Hybrid Loss.

Weighted Logarithmic Loss (WLL). We assign higher penalties to the minority class (default cases):

$$\mathcal{L}_{\text{WLL},i} = -\Big(w^+ y_i \log(y'_i + \epsilon) + w^-(1 - y_i)\log(1 - y'_i + \epsilon) \Big). \tag{4}$$

Dynamic Hard Example Mining. We calculate a dynamic weight ω_i for each sample based on its prediction uncertainty, measured by the gradient norm $g_i = |\partial \mathcal{L}_i / \partial y'_i|$:

$$\omega_i = \frac{(g_i + \epsilon)^{-\alpha}}{\sum_j (g_j + \epsilon)^{-\alpha}}. \tag{5}$$

This mechanism automatically up-weights samples where the model struggles.

Total Objective. The final objective balances regression (for probability smoothness) and classification stability:

$$\mathcal{L}_{\text{total}} = \frac{1}{n} \sum_{i=1}^{n} \omega_i \left[\beta (y'_i - y_i)^2 + (1 - \beta) \mathcal{L}_{\text{WLL},i} \right]. \tag{6}$$

4 Experiment

4.1 Experimental Setup

We first evaluate our method on an industrial credit risk task and additionally verify it on public time series classification datasets.

4.1.1 Industrial Credit Risk Assessment

Dataset. We evaluate FinLangNet on our large-scale proprietary industrial dataset comprising over 7 million users from a real-world financial service platform. The dataset includes 6 evaluation tasks representing overdue status predictions at different future time horizons (denoted as $\tau \in \{1, 2, ..., 6\}$), enabling assessment of both short-term and long-term risk forecasting capabilities. Dataset statistics are provided in Appendix Table 4.

Evaluation Metrics. Following industry standards for credit risk assessment, we employ two primary metrics: (1) the Kolmogorov-Smirnov (KS) statistic (Massey Jr, 1951), which measures the maximum divergence between cumulative distributions of positive and negative classes, directly quantifying discriminative power; and (2) Area Under the Curve (AUC), which evaluates the model's classification ability across all threshold settings.

Baselines. We compare FinLangNet against a diverse set of methods: (1) XGBoost (the current production baseline); (2) deep sequential models including Transformer (Vaswani et al., 2017), Mamba (Gu and Dao, 2023), and TimesNet (Wu et al., 2023); and (3) state-of-the-art LLMs, specifically DeepSeek-V3.2 and GPT-4.1, evaluated in a zero-shot setting to assess general reasoning capabilities. Detailed information in Appendix B.

4.1.2 Public Time Series Benchmark

Dataset. To validate the generalizability of our sequential representation module (SRG), we evaluate on 5 multivariate time series classification (MTSC) tasks from the UEA archive (Bagnall et al., 2018). These datasets span diverse domains including gesture recognition, audio analysis, and medical diagnostics, as detailed in Appendix Table 5.

Evaluation Metrics. Following standard practice for time series classification, we report classification accuracy as the primary metric.

Baselines. We compare against 11 state-of-the-art MTSC methods including distance-based approaches (EDI, DTWI), deep learning models (MLSTM-FCNs (Karim et al., 2019), TapNet (Zhang et al., 2020), ShapeNet (Li et al., 2021)),

Model	$y_1(\tau=1)$		$y_2(\tau=2)$		$y_3(\tau=3)$		$y_4(\tau=4)$		$y_5(\tau=5)$		$y_6(\tau=6)$	
	AUC	KS	AUC	KS	AUC	KS	AUC	KS	AUC	KS	AUC	KS
Tabular & Graph Baselines												
XGBoost	72.78	32.85	75.76	37.42	70.89	30.00	73.04	33.18	68.69	26.96	69.70	28.31
MLP (Goodfellow et al., 2016)	71.97	31.81	74.76	36.14	69.95	28.70	72.00	31.76	67.59	25.38	68.45	26.55
GraphSAGE (Hamilton et al., 2017)	72.19	32.04	75.09	36.49	70.26	29.09	72.37	32.16	68.03	26.00	68.95	27.25
Sequential Baselines												
GRU (Cho et al., 2014)	72.59	32.40	75.68	37.16	70.93	30.05	73.37	33.57	69.06	27.44	70.62	29.75
Transformer (Vaswani et al., 2017)	72.54	32.62	75.95	37.98	70.97	30.12	73.76	34.54	69.30	27.82	71.19	30.67
Mamba (Gu and Dao, 2023)	72.28	32.06	75.66	37.28	70.65	29.67	73.16	33.47	68.79	26.88	70.23	29.02
TimesNet (Wu et al., 2023)	72.49	32.54	75.90	37.98	70.83	29.99	73.48	34.15	69.26	27.73	71.05	30.53
LLM Baselines (Zero-Shot)												
DeepSeek-V3.2 (DeepSeek-AI et al., 2025)	54.80	9.10	55.45	10.50	54.74	7.90	55.94	9.90	54.62	7.70	54.56	7.60
GPT-4.1 (OpenAI, 2025)	55.90	10.85	56.80	12.50	55.15	9.30	56.60	11.60	56.12	10.70	56.05	10.60
FinLangNet (Ours)	73.55	34.08	76.96	39.46	71.92	31.60	74.51	35.67	70.33	29.27	**72.12**	**32.16**
FinLangNet + XGB (Fusion)	**75.11**	**36.55**	**77.76**	**40.73**	**73.04**	**33.27**	**75.32**	**36.81**	**70.39**	**29.31**	71.49	30.97

Table 1: Performance comparison across multiple models and labels. The FinLangNet + XGB represents the deployed fusion strategy. The best results are highlighted in **bold**. All metrics are reported as percentages (%).

feature-based methods (WEASEL-MUSE (Schäfer and Leser, 2017), Rocket (Dempster et al., 2020)), and transformer architectures (TStamp (Zerveas et al., 2021), SVP-T (Zuo et al., 2023)).

4.2 Results on Industrial Assessment

Table 1 presents the comparative evaluation on our large-scale industrial dataset across six temporal prediction tasks (y_1 to y_6). We compare FinLangNet against three categories of baselines: tabular methods, sequential deep learning, and Large Language Models. FinLangNet consistently outperforms all baselines, achieving significant improvements in both AUC and KS metrics. Specifically, compared to the strongest tabular baseline XGBoost, our method yields an average KS improvement of 6.3 pp, demonstrating the superiority of our multi-scale sequential learning approach over static feature engineering. Interestingly, while the fusion of FinLangNet and XGBoost generally achieves the state-of-the-art, it underperforms the standalone FinLangNet on the long-term task y_6. This suggests that XGBoost, which relies on static tabular features, struggles to capture long-range dependencies, thereby introducing noise rather than signal for distant prediction horizons.

Beyond raw performance, scalability is critical for industrial deployment. While some state-of-the-art methods like ShapeFormer (Le et al., 2024) achieve high accuracy on small academic benchmarks, they rely on computationally expensive preprocessing (e.g., shapelet extraction from the entire dataset), making them intractable for our production environment with millions of users. In contrast, FinLangNet is designed for efficiency. Our dual-prompt mechanism processes multi-source sequences without heavy pre-computation, enabling real-time inference. This balance of high predictive accuracy and practical scalability confirms FinLangNet as a viable solution for large-scale industrial credit risk systems.

4.3 Results on UEA Time Series Classification

To verify the generalization capability of our approach beyond financial domains, we evaluate the stand-alone SRG module on public UEA benchmarks. As illustrated in Figure 2, our method achieves competitive or state-of-the-art performance across diverse datasets ranging from medical diagnostics such as AtrialFibrillation to motion recognition like BasicMotions. Even without the full FinLangNet architecture by utilizing only the SRG module, our approach consistently outperforms transformer-based baselines such as TStamp and SVP-T. Detailed numerical comparisons are provided in Appendix Table 9.

These results highlight the versatility of our dual-prompt mechanism. While specifically designed to handle the heterogeneity of financial data, the SRG module proves effective in general multivariate time series classification tasks. Notably, it achieves this performance while maintaining the computational efficiency necessary for industrial deployment unlike heavy academic models like ShapeFormer. This confirms that the architectural innovations in FinLangNet, specifically the multi-granularity prompts, capture fundamental temporal dependencies applicable across various domains.

Length Setup	y_1		y_2		y_3		y_4		y_5		y_6		Avg KS
	AUC	KS	AUC	KS	AUC	KS	AUC	KS	AUC	KS	AUC	KS	(%)
1.00 × Length (Full)	**73.55**	**34.08**	**76.96**	**39.46**	**71.92**	**31.60**	**74.51**	**35.67**	**70.33**	**29.27**	**72.12**	**32.16**	**33.71**
0.50 × Length	73.45	34.00	76.80	39.34	71.80	31.55	74.45	35.61	70.22	29.14	72.07	32.11	33.62
0.25 × Length	73.36	33.82	76.77	39.18	71.72	31.30	74.37	35.54	70.14	29.07	71.99	32.00	33.49

Table 2: Performance robustness across varying historical sequence lengths. The model maintains high stability even when the input history is significantly truncated.

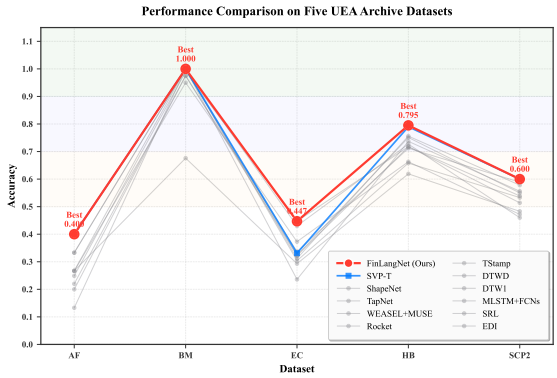

Figure 2: Accuracy comparison on selected UEA datasets. Our SRG module demonstrates consistent superiority or competitiveness against SOTA benchmarks.

Configuration	KS (%)	AUC (%)
Baseline (No Prompts)	32.62	72.54
Prompt Contributions		
+ Feature Prompt ($\widetilde{\phi}_c$)	32.79	72.65
+ Source Prompt (P_s)	32.99	72.78
+ Both Prompts (Ours)	**34.08**	**73.55**
Module Importance		
w/o Sequential Model (O_{ns})	12.53	58.62
w/o Non-sequential Model (O_m)	32.75	72.66

Table 3: Ablation analysis of different modules on task $y_1(\tau = 1)$. The proposed dual-prompt mechanism ($\widetilde{\phi}_c + P_s$) combined with sequential modeling achieves the best performance.

5 Ablation Study

To understand the contribution of modules in FinLangNet, we conduct comprehensive ablation studies on the industrial credit risk dataset. All experiments strictly maintain consistent hyperparameters, varying only the component under evaluation.

5.1 Component Analysis

Table 3 dissects the impact of different architectural choices. The results validate the efficacy of our dual-prompt mechanism where combining the feature-granularity prompt $\widetilde{\phi}_c$ and the data-source prompt P_s yields the highest performance of 34.08% KS. This outcome confirms that modeling both local feature nuances and global cross-source interactions provides complementary benefits. Crucially, removing the sequential module O_{ns} leads to a catastrophic performance drop where the KS decreases by over 20%, demonstrating that temporal behavioral dynamics are the dominant predictor of credit risk while static non-sequential features serve primarily as a supplementary signal.

5.2 Impact of Sequence Length

We further evaluate the model's robustness by truncating the historical sequence length while keeping non-sequential features fixed as shown in Table 2. The performance remains remarkably stable even when the input sequence is reduced to only 25% of its original length, with the average KS score dropping marginally from 33.71% to 33.49%. This results indicates that recent short-term behavioral patterns encode the most critical risk signals, which allows FinLangNet to generalize effectively to users with limited transaction histories and serves as a vital characteristic for solving the cold-start problem in production environments.

5.3 Online A/B Test

In real-world financial risk management, regulatory compliance bans the use of black-box models for direct credit decisioning. To bridge the gap between deep learning performance and industrial explainability, we designed a fusion framework. We deployed FinLangNet to abstract heterogeneous data streams (e.g., inquiry records, account ledgers, and behavioral logs) into a scalar language-risk subscore, s_{lang}. This score is then integrated into our interpretability-centric XGBoost system as a dense feature, effectively enhancing the model's discriminative power without violating risk governance protocols. Theoretically complex but operationally efficient, the system runs on L20 GPU clusters, processing 100 QPS with sub-100ms latency.

Online performance analysis detailed in Appendix D and Figure 4 demonstrates significant risk

reduction. In a controlled A/B test at a representative 60% approval threshold, the proposed framework reduced the default rate from 9.1% in the benchmark to 8.2%, which represents a 9.9% relative improvement. This confirms that FinLangNet effectively discriminates risk by systematically rejecting high-risk applicants that traditional models miss. Further analysis reveals that combining sequential representations with domain features yields a 6.3 pp improvement in KS and a 12.3% reduction in expected losses, thus validating the commercial value of the hybrid framework.

6 Related Work

Credit risk prediction is a crucial task in the financial sector, typically focused on estimating the likelihood of borrower default over a specific time period. Credit scores, such as FICO (Maiden and Maiden, 2024; Lei et al., 2025a), are widely used evaluation tools (Jensen, 1992; Bücker et al., 2022), generated by algorithms that analyze various user-related data to assess a borrower's creditworthiness (Zhang et al., 2025). Extensive research has explored the application of machine learning techniques for credit risk prediction, including decision-tree-based methods like XGBoost (He et al., 2018), graph models such as GraphSAGE (Balmaseda et al., 2023), ChebConv (Liu, 2022), and their combinations (Fein-Ashley et al., 2024). While deep learning methods are often considered to offer enhanced modeling capabilities, existing studies have found that XGBoost generally outperforms deep learning approaches in this domain (Xu et al., 2021).

The processing of irregular, multi-source financial data is a critical challenge in credit risk prediction. Such data are typically presented in tabular form and often involve high-dimensional features, necessitating effective feature selection techniques. Several methods have been proposed to improve performance, including filter methods (Janane et al., 2023), wrapper methods (Ahadzadeh et al., 2023), and embedded methods (Raghu et al., 2023), which enhance both model accuracy (Xu et al., 2024; Ha et al., 2019; Li et al., 2020) and interpretability (Ma et al., 2018; Xu et al., 2021). Deep learning approaches have also been applied to handle these diverse data types (Gorishniy et al., 2021; Borisov et al., 2022), but they do not consistently outperform XGBoost models (Gorishniy et al., 2021) when working with tabular data.

From data perspective, another approach is to treat financial information as sequential data for processing. Structured data based on temporal sequences, such as transaction records or historical behaviors, can be represented in the standard time-series format, which encompass a wide range of methodologies. Models like EDI, DTWI, and DTWD (Bagnall et al., 2018) rely on calculating distances reflective of temporal warping or deviations in time sequences. MLSTM-FCNs (Karim et al., 2019) combine LSTM and CNN layers for feature generation, while WEASEL-MUSE (Schäfer and Leser, 2017) transforms series into symbolic representations. In the domain of credit risk, time-series models include approaches from both statistical and machine learning (El-Qadi et al., 2022), as well as deep learning methods (Forough and Momtazi, 2022; Ala'raj et al., 2021; Wang and Xiao, 2022; Liang et al., 2023) such as LSTM (Yu et al., 2019) and Transformer (Vaswani et al., 2017; Lei et al., 2025b). However, existing methods are predominantly built on preprocessed, well-structured financial time-series data, which limits their ability to handle irregular, multi-source records. Additionally, most approaches rely on relatively plain model architectures.

7 Conclusion

In this work, we present FinLangNet, a novel framework that addresses critical challenges in industrial credit risk assessment by treating it as a multi-scale behavioral prediction task. Our approach integrates heterogeneous data streams through a dual-architecture design that combines DeepFM for static features and the Sequential Representation Generation module with a dual-prompt mechanism for temporal data. Extensive experiments demonstrate FinLangNet's effectiveness with a 6.3 pp improvement in KS metric over XGBoost and a 9.9% reduction in relative bad debt rate on industrial datasets while achieving state-of-the-art performance on UEA time series benchmarks. Successfully deployed in the international finance system of a leading ride-hailing platform serving millions of daily transactions, FinLangNet validates the feasibility of combining academic innovation with industrial practicality in high-stakes financial applications.

References

Behrouz Ahadzadeh, Moloud Abdar, Fatemeh Safara, Abbas Khosravi, Mohammad Bagher Menhaj, and Ponnuthurai Nagaratnam Suganthan. 2023. Sfe: A simple, fast and efficient feature selection algorithm for high-dimensional data. *IEEE Transactions on Evolutionary Computation*.

Maher Ala'raj, Maysam F Abbod, and Munir Majdalawieh. 2021. Modelling customers credit card behaviour using bidirectional lstm neural networks. *Journal of Big Data*, 8(1):69.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. 2018. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*.

Vicente Balmaseda, María Coronado, and Gonzalo de Cadenas-Santiago. 2023. Predicting systemic risk in financial systems using deep graph learning. *Intelligent Systems with Applications*, 19:200240.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Michael Bücker, Gero Szepannek, Alicja Gosiewska, and Przemyslaw Biecek. 2022. Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, 73(1):70–90.

Dawei Cheng, Zhibin Niu, and Yiyi Zhang. 2020. Contagious chain risk rating for networked-guarantee loans. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2715–2723.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Angus Dempster, François Petitjean, and Geoffrey I Webb. 2020. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ayoub El-Qadi, Maria Trocan, Thomas Frossard, and Natalia Díaz-Rodríguez. 2022. Credit risk scoring forecasting using a time series approach. In *Physical Sciences Forum*, volume 5, page 16. MDPI.

Gianluca Elia, Valeria Stefanelli, and Greta Benedetta Ferilli. 2023. Investigating the role of fintech in the banking industry: what do we know? *European Journal of Innovation Management*, 26(5):1365–1393.

Jacob Fein-Ashley, Tian Ye, Sachini Wickramasinghe, Bingyi Zhang, Rajgopal Kannan, and Viktor Prasanna. 2024. A single graph convolution is all

you need: Efficient grayscale image classification. *arXiv preprint arXiv:2402.00564*.

Javad Forough and Saeedeh Momtazi. 2022. Sequential credit card fraud detection: A joint deep neural network and probabilistic graphical model approach. *Expert Systems*, 39(1):e12795.

Sergio Genovesi, Julia Maria Mönig, Anna Schmitz, Maximilian Poretschkin, Maram Akila, Manoj Kahdan, Romina Kleiner, Lena Krieger, and Alexander Zimmermann. 2023. Standardizing fairness-evaluation procedures: interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans. *AI and Ethics*, pages 1–17.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*.

Van-Sang Ha, Dang-Nhac Lu, Gyoo Seok Choi, Ha-Nam Nguyen, and Byeongnam Yoon. 2019. Improving credit risk prediction in online peer-to-peer (p2p) lending using feature selection with deep learning. In *2019 21st International Conference on Advanced Communication Technology (ICACT)*, pages 511–515. IEEE.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30.

Hongliang He, Wenyu Zhang, and Shuai Zhang. 2018. A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, 98:105–117.

Fatima Zahra Janane, Tayeb Ouaderhman, and Hasna Chamlal. 2023. A filter feature selection for high-dimensional data. *Journal of Algorithms & Computational Technology*, 17:17483026231184171.

Herbert L Jensen. 1992. Using neural networks for credit scoring. *Managerial finance*, 18(6):15–26.

Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. 2019. Multivariate lstm-fcns for time series classification. *Neural networks*, 116:237–245.

Xuan-May Le, Ling Luo, Uwe Aickelin, and Minh-Tuan Tran. 2024. Shapeformer: Shapelet transformer for multivariate time series classification. *arXiv preprint arXiv:2405.14608*.

Yu Lei, Zixuan Wang, Chu Liu, and Tongyao Wang. 2025a. Zigong 1.0: A large language model for financial credit. *arXiv preprint arXiv:2502.16159*.

Yu Lei, Jiayang Zhao, Yilei Zhao, Zhaoqi Zhang, Linyou Cai, Qianlong Xie, and Xingxing Wang. 2025b. Generative large-scale pre-trained models for automated ad bidding optimization. *arXiv preprint arXiv:2508.02002*.

Guozhong Li, Byron Choi, Jianliang Xu, Sourav S Bhowmick, Kwok-Pan Chun, and Grace Lai-Hung Wong. 2021. Shapenet: A shapelet-neural network approach for multivariate time series classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8375–8383.

Wei Li, Shuai Ding, Hao Wang, Yi Chen, and Shanlin Yang. 2020. Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in china. *World Wide Web*, 23:23–45.

Yixuan Li, Charalampos Stasinakis, and Wee Meng Yeo. 2022. A hybrid xgboost-mlp model for credit risk assessment on digital supply chain finance. *Forecasting*, 4(1):184–207.

Yancheng Liang, Jiajie Zhang, Hui Li, Xiaochen Liu, Yi Hu, Yong Wu, Jinyao Zhang, Yongyan Liu, and Yi Wu. 2023. Derisk: An effective deep learning framework for credit risk prediction over real-world financial data. *arXiv preprint arXiv:2308.03704*.

Xinyan Liu. 2022. Fast recommender system combining global and local information: Construction of large-scale commodity information recommendation system. In *2022 2nd International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR)*, pages 166–169. IEEE.

Tian Lu and Yingjie Zhang. 2023. Profit vs. equality? the case of financial risk assessment and a new perspective on alternative data. *MIS Quarterly*, 47(4).

Xiaojun Ma, Jinglan Sha, Dehua Wang, Yuanbo Yu, Qian Yang, and Xueqi Niu. 2018. Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31:24–39.

Stephen E Maiden and Stephen E Maiden. 2024. Fico score. *Darden Business Publishing Cases*, pages 1–11.

Frank J Massey Jr. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Majid Niazkar, Andrea Menapace, Bruno Brentan, Reza Piraei, David Jimenez, Pranav Dhawan, and Maurizio Righetti. 2024. Applications of xgboost in water resources engineering: A systematic literature review (dec 2018–may 2023). *Environmental Modelling & Software*, page 105971.

OpenAI. 2025. Introducing openai gpt-4.1.

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. 2023. Sequential multi-dimensional self-supervised learning for clinical time series. In *International Conference on Machine Learning*, pages 28531–28548. PMLR.

Patrick Schäfer and Ulf Leser. 2017. Multivariate time series classification with weasel+ muse. *arXiv preprint arXiv:1711.11343*.

Yu Song, Yuyan Wang, Xin Ye, Russell Zaretzki, and Chuanren Liu. 2023. Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme. *Information Sciences*, 629:599–617.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chongren Wang and Zhuoyi Xiao. 2022. A deep learning approach for credit scoring using feature embedded transformer. *Applied Sciences*, 12(21):10995.

Haixu Wu, Zonghan Xu, Yuxuan Liu, Yihan Wu, Jing Lin, Li Zeng, and Xing Xie. 2023. Timesnet: Temporal 2d-variation modeling for general time series analysis. *International Conference on Learning Representations (ICLR)*.

Jinxin Xu, Han Wang, Yuqiang Zhong, Lichen Qin, and Qishuo Cheng. 2024. Predict and optimize financial services risk using ai-driven technology. *Academic Journal of Science and Technology*, 10(1):299–304.

Junhui Xu, Zekai Lu, and Ying Xie. 2021. Loan default prediction of chinese p2p market: a machine learning methodology. *Scientific Reports*, 11(1):18759.

Gang Xue, Shifeng Liu, Long Ren, and Daqing Gong. 2024. Risk assessment of utility tunnels through risk interaction-based deep learning. *Reliability Engineering & System Safety*, 241:109626.

Yuantao Yao, Minghan Yang, Jianye Wang, and Min Xie. 2022. Multivariate time-series prediction in industrial processes via a deep hybrid network under data uncertainty. *IEEE Transactions on Industrial Informatics*, 19(2):1977–1987.

Samuel Yousefi and Babak Mohamadpour Tosarkani. 2022. The adoption of new technologies for sustainable risk management in logistics planning: A sequential dynamic approach. *Computers & Industrial Engineering*, 173:108627.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124.

Junru Zhang, Lang Feng, Xu Guo, Yuhan Wu, Yabo Dong, and Duanqing Xu. 2025. Timemaster: Training time-series multimodal llms to reason via reinforcement learning. *arXiv preprint arXiv:2506.13705*.

Xuchao Zhang, Yifeng Gao, Jessica Lin, and Chang-Tien Lu. 2020. Tapnet: Multivariate time series classification with attentional prototypical network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6845–6852.

Yang Zhao, John W Goodell, Yong Wang, and Mohammad Zoynul Abedin. 2023. Fintech, macroprudential policies and bank risk: Evidence from china. *International Review of Financial Analysis*, 87:102648.

Rundong Zuo, Guozhong Li, Byron Choi, Sourav S Bhowmick, Daphne Ngar-yin Mah, and Grace LH Wong. 2023. Svp-t: a shape-level variable-position transformer for multivariate time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11497–11505.

Appendix

The Appendix is organized as follows:

A Data Details and Preprocessing
details the specifications of the large-scale Industrial Credit Risk Dataset (including non-sequential and sequential splits) and the public UEA benchmarks.

B Hyperparameter Settings
Lists the detailed training configurations (Table 6) and FinLangNet architecture specifications (Table 7).

C Extended UEA Benchmark Results
Validation of generalization capability on 5 UEA time-series datasets compared with state-of-the-art baselines.

D Real-World Deployment Analysis
An industrial case study, comparing FinLangNet with the production XGBoost model on thresholds, risk distribution, and operational impact.

A Data Details and Preprocessing

In this section, we provide comprehensive specifications for the datasets utilized in our experiments. We evaluate our method on two distinct types of data: (1) a large-scale industrial credit risk dataset, and (2) public benchmarks from the UEA Multivariate Time Series Archive.

A.1 Industrial Credit Risk Dataset

We collected a real-world dataset from a leading international financial platform to evaluate credit risk prediction. The data originates from real business lines and is presented in tabular form. As illustrated in Figure 3, we reorganized the raw multi-source data into static and sequential components based on their temporal nature.

Data Composition. The input X is derived from three main parts: i) **Basic Information**: Utilized as non-sequential data (m), containing invariant user attributes. ii) **Credit Report**: Further split into Inquiry Records and Account Records. The reports are collected during the credit underwriting stage, with explicit user authorization granted to the platform for querying licensed third-party credit bureaus. These paid bureau inquiries provide detailed credit information at the individual record level. iii) **Loan Behavior**: Logs of borrowing and repayment. The latter two parts constitute the sequential data (z). We unified these records in strict chronological order. Statistical details are summarized in Table 4. The dataset involves approximately 3.6 million users, split temporally into 3 million for training and 0.6 million for testing.

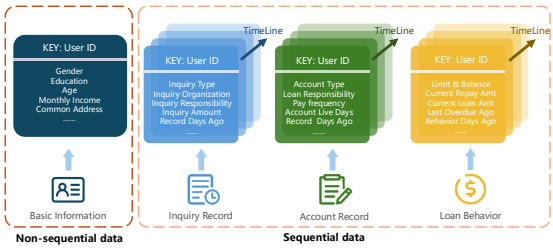

Figure 3: Structure of the Industrial Credit Risk Dataset. The input is categorized into Non-Sequential static data (m) and Sequential time-dependent data (z).

A.2 UEA Public Benchmarks

To verify the generalization capability of our model, we also evaluate it on 5 representative multivariate time series classification datasets from the UEA Archive (Bagnall et al., 2018). These datasets cover diverse domains including medical monitoring and human motion detection. The details are listed in Table 5.

DeepFM Formulation In the Non-Sequential Module, we utilize DeepFM to capture interactions within the static feature vector $m \in \mathbb{R}^M$. The **FM component** models second-order feature interactions via inner products of latent vectors V:

$$y_{\text{FM}} = \langle w, m \rangle + \sum_{j_1=1}^{M} \sum_{j_2=j_1+1}^{M} \langle V_{j_1}, V_{j_2} \rangle m_{j_1} m_{j_2}. \tag{7}$$

Simultaneously, the **Deep component** models high-order interactions using a Multi-Layer Perceptron (MLP). The embeddings are fed into the network as:

$$y_{\text{DNN}} = \text{MLP}\left(\frac{1}{2}\left[\left(\sum_{i=1}^{M} m_i V_i\right)^2 - \sum_{i=1}^{M} \left(m_i^2 V_i^2\right)\right]\right). \tag{8}$$

The final output combines both components: $O_m = \sigma(y_{\text{FM}} + y_{\text{DNN}})$.

B Hyperparameter Settings

We provide detailed hyperparameter configurations used in our experiments to ensure reproducibility. Table 6 lists the optimization and training parameters for our proposed FinLangNet. Table 7 details the specific architecture specifications of our model. Additionally, Table 8 provides the specific versions and inference configurations for the Large Language Model (LLM) baselines.

Data Source	Channel	Seq. Len.	Dataset Split	Description
Basic Information	11	–	(3M, 0, 0.6M)	Static Personal Profile
Credit Report: Inquiry	5	120		External Credit Queries
Credit Report: Account	29	200		History of Credit Lines
Loan Behavior	241	26		Repayment & Borrowing Logs

Table 4: Statistics of the Industrial Credit Risk Dataset. **Channel** denotes the feature dimension (C_s), and **Seq. Len.** denotes the maximum sequence length (T_s).

Dataset	Train Cases	Test Cases	Dimensions	Length	Classes
AtrialFibrillation (AF)	15	15	2	640	3
BasicMotions (BM)	40	40	4	100	4
EthanolConcentration (EC)	261	263	3	1751	4
Heartbeat (HB)	204	105	61	495	2
SelfRegulationSCP2 (SCP2)	200	180	7	1152	2

Table 5: Details of the utilized UEA Multivariate Time Series Classification datasets.

C Extended UEA Benchmark Results

To validate the generalization capability of FinLangNet beyond financial applications, we conducted comprehensive experiments on the UEA multivariate time series archive. Tables 9 present detailed comparisons with state-of-the-art time series classification methods.

D Real-World Deployment Analysis

To demonstrate the practical advantages of FinLangNet in industrial settings, we present a detailed case study comparing our approach with the production XGBoost model currently deployed at real-world . This analysis focuses on real-world performance metrics and operational considerations critical for risk management systems.

D.1 Experimental Setup

We compared two modeling paradigms:

- **XGBoost (Baseline):** The production model utilizing 500+ manually engineered features derived from domain expertise, optimized over years of deployment.
- **FinLangNet:** Our proposed model processing raw sequential data without manual feature engineering, trained with multi-task learning across seven risk-related objectives.

For fair comparison, both models were evaluated on the common primary target y_1 (30-day delinquency, $\tau = 1$), which represents the most business-critical risk indicator. Performance was assessed on a held-out one-month test window following the training period.

D.2 Threshold-Based Performance Analysis

Risk control systems require careful threshold selection to balance between risk exposure (false negatives) and customer experience (false positives). Figure 5 illustrates the precision-recall trade-offs at various decision thresholds. Key observations include:

- **Low Thresholds (0.0–0.2):** XGBoost achieves higher recall but suffers from poor precision, creating operational challenges with excessive false positives.
- **Operational Range (0.2–0.4):** FinLangNet maintains balanced performance with significantly higher precision while preserving competitive recall, reducing manual review costs.
- **High Thresholds (0.4+):** Both models converge in performance, though FinLangNet maintains a slight precision advantage.

D.3 Risk Distribution Analysis

Figure 6 reveals fundamental differences in how the models discriminate risk.

- **Risk Separation:** FinLangNet produces clearer separation between defaulters and non-defaulters, especially in the high-risk segment (predicted scores > 0.6).
- **Score Calibration:** While XGBoost shows some score clustering around certain values (likely due to dominant features), FinLangNet provides more continuous risk scoring.

Parameter	Value	Description
Optimizer	AdamW	Optimizer with weight decay for better generalization
Learning Rate	5e-4	Initial learning rate with cosine annealing schedule
Betas	(0.9, 0.999)	Exponential decay rates for Adam moment estimates
Epsilon	1e-08	Numerical stability constant for Adam
Weight Decay	0.01	L2 regularization coefficient
Loss Weight α	0.5	Weight for auxiliary task loss in multi-task learning
Loss Weight β	0.5	Balance between MSE (continuous) and WLL (categorical)
Focal Weight γ	1.0	Focal loss parameter for hard sample mining
Training Epochs	12	Total training iterations over full dataset
Batch Size	512	Number of samples per gradient update
Gradient Clipping	1.0	Maximum gradient norm for stability
Dropout Rate	0.2	Dropout probability for regularization
Early Stopping Patience	3	Epochs to wait before stopping if no improvement

Table 6: Training hyperparameters and optimization settings for FinLangNet.

Component	Configuration	Description
Sequential Branch (SRG Module)		
Embedding Dimension	512	Dimension of token embeddings
Transformer Layers	10	Number of self-attention layers
Attention Heads	16	Multi-head attention configuration
FFN Hidden Size	512	Feed-forward network intermediate dimension
Position Encoding	Learnable	Trainable positional embeddings
Max Sequence Length	200	Maximum tokens per sequence
Prompt Pool Size	16	Number of learnable soft prompts
Prompt Length	10	Tokens per prompt template
Non-Sequential Branch (DeepFM)		
FM Embedding Size	16	Factorization machine latent dimension
DNN Hidden Layers	[512, 128, 64]	Deep network architecture
Activation Function	ReLU	Non-linearity for hidden layers
Batch Normalization	True	Applied after each hidden layer
Fusion Layer		
Fusion Method	Attention	Cross-modal attention mechanism
Output Dimension	128	Final representation size
Temperature τ	0.07	Contrastive learning temperature

Table 7: FinLangNet architecture specifications.

Model Family	Configuration	Description
GPT-4 Series		
Model Version	gpt-4.1-2025-04-14	Specific snapshot API version utilized for experiments
Inference Temperature	0.0	Set to zero for deterministic and reproducible evaluation
Top-p	1.0	Nucleus sampling probability
Max Output Tokens	512	Limit for reasoning chain and classification label
DeepSeek Series		
Model Version	deepseek-v3-2-251201	Specific snapshot version utilized for experiments
Inference Temperature	0.0	deterministic generation setting
Repetition Penalty	1.1	Penalty to prevent circular reasoning in risk analysis
Context Window	128k	Full context window utilized for long behavior logs
Common Zero-Shot Setup		
System Prompt	Expert Role	"You are a senior credit risk expert..."
Input Serialization	JSON	Structured sequence representation for financial data

Table 8: Configuration for Large Language Model (LLM) baselines. We utilize specific snapshot versions to ensure experimental consistency.

Data	EDI	DTWI	DTWD	MLSTM-FCNs	WEASEL+MUSE	SRL	TapNet	ShapeNet	Rocket	TStamp	SVP-T	FinLangNet
AF	0.267	0.267	0.220	0.267	0.333	0.133	0.333	**0.400**	0.249	0.200	**0.400**	**0.400**
BM	0.676	**1.000**	0.975	0.950	**1.000**	**1.000**	**1.000**	**1.000**	0.990	0.975	**1.000**	**1.000**
EC	0.293	0.304	0.323	0.373	0.430	0.236	0.323	0.312	**0.447**	0.337	0.331	**0.447**
HB	0.619	0.658	0.717	0.663	0.727	0.737	0.751	0.756	0.718	0.712	0.790	**0.795**
SCP2	0.483	0.533	0.539	0.472	0.460	0.556	0.550	0.578	0.514	0.589	**0.600**	**0.600**

Table 9: Accuracies on Five Datasets of the UEA Archive

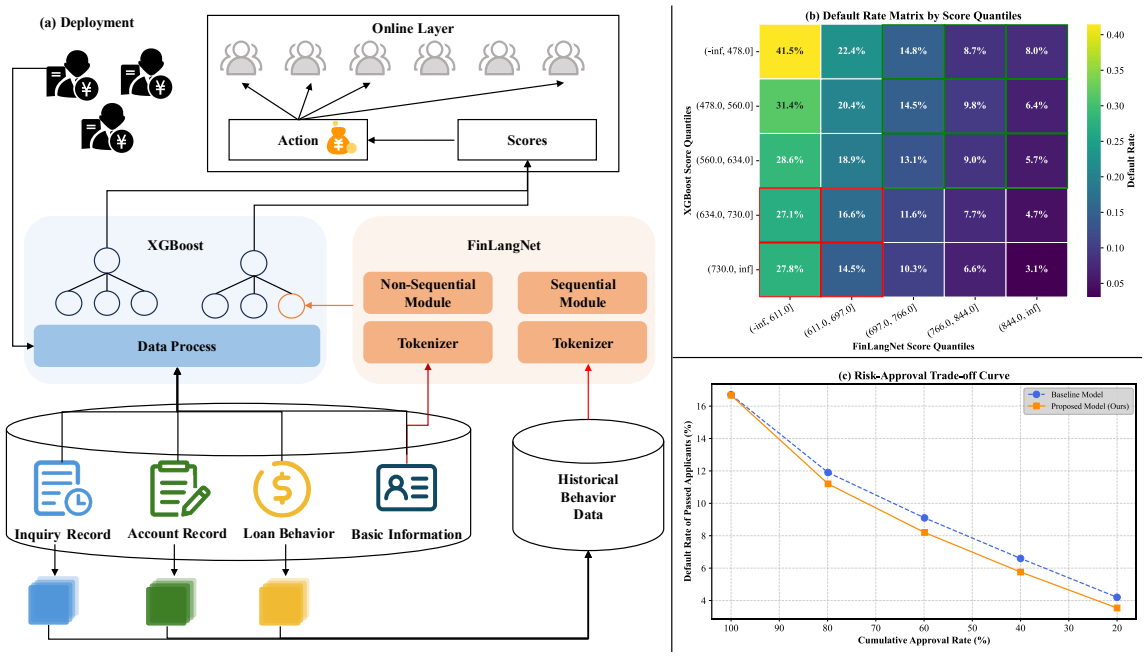

Figure 4: Overview of Deployment. (a) Online deployment architecture and data flow. (b) Default-rate matrix by score quantiles: The red-circled region indicates applicants rejected by FinLangNet but approved by the benchmark (risk reduction opportunities), while the green region shows potential viable customers. (c) Risk–approval trade-off demonstrating consistent default rate reduction across varying approval thresholds.

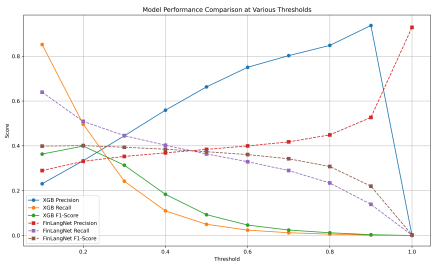

Figure 5: Performance comparison at various risk thresholds for $y_1(\tau = 1)$ prediction. FinLangNet demonstrates superior precision at operational thresholds (0.2–0.4) commonly used in production.

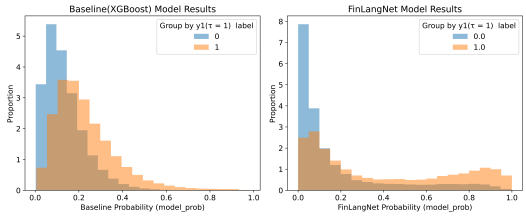

Figure 6: Distribution of predicted risk scores versus actual labels. FinLangNet exhibits better separation between risk classes, particularly in the high-risk segment (scores > 0.6).

- **False Positive Distribution:** FinLangNet's false positives tend to cluster in the moderate-risk range (0.3–0.5), making them easier to identify through secondary screening.

D.4 Operational Impact Analysis

Recognizing the complementary strengths of both approaches, we developed a hybrid strategy that incorporates FinLangNet's representations as additional features in the XGBoost framework. This integration leverages FinLangNet's automated feature learning from raw sequential data alongside XGBoost's domain-specific features. The hybrid model achieved a **6.3 pp improvement in KS metric** over standalone XGBoost. To assess real-world deployment benefits, we analyzed swap sets—cases where the two models disagree on risk classification—and tracked their actual delinquency outcomes:

- **High-Risk Captures:** Among users classified as high-risk only by FinLangNet, 68% showed delinquent behavior within 60 days, validating its superior risk detection.
- **False Positive Reduction:** FinLangNet correctly identified 42% of XGBoost's false positives as low-risk, potentially reducing unnecessary credit restrictions.
- **Portfolio Performance:** Applying FinLangNet's risk scores would reduce the portfolio default rate by an estimated 12.3% while maintaining the same approval rate.