Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Deep Learning Architectures for Credit Risk Assessment

**Balogun David Taiwo**

balogundavi98@gmail.com

University of Lagos

**Research Article**

**Keywords:** Transformer models, Gated Recurrent Unit, Deep Learning, sequence based LSTM models, Long Short Term Memory

**Posted Date:** January 19th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-8621356/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Abstract

This study investigates the effectiveness of advanced deep learning architectures including Long Short Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), and Transformer models for predicting credit default using sequential borrower behavior data. Traditional credit scoring systems often rely on static features and overlook temporal dependencies inherent in financial activity. To address this limitation, we develop and evaluate a suite of deep learning models that learn from time ordered transaction histories, payment behaviors, and utilization patterns. In addtion, we explore ensemble strategies that combine neural representations with classical statistical approaches, such as logistic regression and gradient boosting, to improve robustness and interpretability.

Comprehensive experiments are conducted on both publicly available datasets (e.g., LendingClub) and proprietary microfinance loan portfolios. Evaluation metrics include AUC-ROC, log loss, and calibration error. Results show that Transformer-based models consistently outperform both traditional models and recurrent neural networks in terms of predictive accuracy and reliability. We also observe that ensemble models offer a favorable balance between performance and explainability, making them more suitable for production environments in regulated industries.

Our findings underscore the transformative potential of deep learning in credit risk assessment, especially when deployed with fairness aware practices and explainability tools. This research contributes to the growing body of work advocating for AI driven credit scoring frameworks that are not only data driven and accurate but also operationally viable and ethically responsible.


# 1. Introduction

In the modern financial landscape, the ability to accurately assess credit risk is not just a matter of profitability it is a necessity for responsible lending and financial inclusion. As digital banking and alternative credit services expand globally, lenders are increasingly relying on data driven approaches to evaluate a borrower’s likelihood of default. Yet, many traditional credit scoring systems remain rooted in static features and linear models, limiting their ability to capture complex behavioral patterns over time.

This research explores a new frontier in credit risk assessment, the use of deep learning architectures to model sequential borrower behavior. Specifically, we investigate how recurrent neural networks namely Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) models as well as the more recently developed Transformer models, can be used to improve the prediction of loan default. These models are uniquely capable of learning temporal dependencies and nonlinear interactions from time stamped financial data, such as payment histories, transaction volumes, and credit utilization trends.

We benchmark the performance of these architectures against well established traditional models, including logistic regression and gradient boosting, using both public datasets (e.g., LendingClub) and proprietary data from a microfinance institution. Beyond model accuracy, we also explore ensemble

strategies that integrate deep learning with classical methods to enhance reliability, robustness, and interpretability key requirements for deployment in real world, regulated financial systems.

Our findings demonstrate that Transformer based models not only outperform other deep learning methods but also offer greater resilience under varying data conditions. Importantly, we show that ensemble approaches combining neural networks with simpler models can deliver high performance while remaining transparent enough for institutional use.

This paper aims to bridge the gap between cutting edge machine learning research and practical credit scoring applications. By doing so, we contribute to a growing movement toward more inclusive, adaptive, and fair financial technologies.

## 2. Related Work

Credit risk assessment has long been a critical area of focus in finance, attracting extensive research from both the academic and industry communities. Traditional credit scoring methods typically rely on statistical models, most notably logistic regression, decision trees, and scorecard based systems. These approaches use static borrower attributes such as income, employment status, credit history, and outstanding debt to predict the likelihood of default. While these models are interpretable and relatively easy to deploy, they often fall short in capturing complex, time-dependent borrower behaviors, such as changing spending habits, fluctuating balances, or delayed payments over time.

In response to these limitations, machine learning techniques have been increasingly applied to credit risk modeling. Tree-based models such as Random Forest and XGBoost have become popular due to their high performance on structured tabular data. Research by Lessmann et al. (2015) showed that ensemble models often outperform single statistical classifiers in credit scoring tasks, especially in terms of predictive accuracy and stability. XGBoost, in particular, has been widely adopted by financial institutions for its ability to handle missing data, nonlinearity, and feature interactions efficiently.

Most of these models treat each loan application or customer profile as a flat vector of features essentially ignoring the sequential nature of borrower behavior. This is a significant shortcoming, especially in the context of digital lending platforms and open banking environments, where time series financial data is readily available. To address this, researchers have turned to deep learning, particularly recurrent neural networks (RNNs) like Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. These models are specifically designed to process sequential data and learn long term dependencies, making them suitable for modeling monthly repayments, cash flow patterns, and credit utilization trajectories. Jurgovsky et al. (2018), for example, demonstrated the effectiveness of sequence based LSTM models in detecting credit card fraud a related but similarly structured risk prediction task.

More recently, Transformer models originally developed for natural language processing have been introduced as a powerful alternative to RNNs. Unlike LSTMs or GRUs, which process sequences step by step, Transformers use self-attention mechanisms that allow them to capture relationships between all

elements in a sequence simultaneously. This not only improves performance on longer sequences but also enables parallel training, making Transformers more scalable. In financial domains, applications of Transformers remain nascent but promising. For instance, the TabTransformer (Huang et al., 2020) architecture has been adapted to tabular financial data and has shown state-of-the-art performance on various classification tasks, including fraud detection and credit scoring.

In addition to exploring model performance, several studies have investigated how to enhance the fairness and interpretability of AI models in finance. Post-hoc explanation tools such as SHAP (SHapley Additive exPlanations) have become popular for quantifying feature importance in complex models. These tools are particularly valuable in regulated environments, where explainability and auditability are crucial. Some studies have also proposed hybrid models combining deep learning with traditional scoring frameworks to maintain predictive performance while satisfying regulatory constraints.

Despite these advancements, there remains a notable gap in the literature around the integration of deep learning into production-level credit risk systems. Most academic studies either focus on benchmarking models on public datasets or fail to address deployment challenges such as fairness, robustness, and model monitoring. Our work contributes to this space by not only comparing deep learning models (LSTM, GRU, Transformer) on real-world credit data, but also by examining their integration with ensemble methods and discussing practical deployment considerations, including fairness audits and interpretability.

In summary, this study builds on prior work in credit scoring by bridging the strengths of modern deep learning techniques with the operational needs of financial institutions. It extends the existing literature through a rigorous comparative analysis, grounded in both theoretical understanding and practical applicability.

# 3.Methodology

This section outlines the technical approach taken to evaluate the performance of deep learning architectures for credit default prediction. Our methodology encompasses problem formulation, model design, training strategies, and ensemble integration.

# 3.1 Problem Formulation

Credit default prediction is framed as a supervised binary classification task. Given a time-ordered sequence of borrower activity such as monthly repayments, balance utilization, late payments, and new credit inquiries the goal is to predict whether the borrower will default on their loan within a fixed future window (e.g., 90 days).

# 3.2 Data Preprocessing

The input sequences are structured at the borrower month level, with each time step representing aggregated borrower behavior during that month. Missing values are imputed using forward fill or

domain-specific defaults (e.g., zero balances). Categorical variables (e.g., loan type, employment status) are embedded, while continuous features are normalized per borrower. Sequences are zero-padded to a maximum length for uniform batch processing.

## 3.3 Model Architectures

We implemented and evaluated three categories of models:

- LSTM and GRU: Recurrent neural networks capable of capturing sequential dependencies. Each model consists of two hidden layers, followed by a fully connected layer with sigmoid activation for binary classification.
- Transformer: Based on the attention-is all you need framework, this model uses positional embeddings and multi head self attention to capture global dependencies. A feed forward network and dropout layers enhance regularization.
- Traditional Baselines: Logistic regression and XGBoost models are trained using aggregated borrower features across the full sequence (e.g., max delinquency, mean balance usage).

## 3.4 Ensemble Approach

To balance performance and interpretability, we use stacking ensembles where the outputs of the deep learning models are treated as meta features, combined using logistic regression or gradient boosting. This allows for confidence calibration and compatibility with traditional scorecards.

## 3.5 Training Details

All models were implemented using TensorFlow and PyTorch. Hyperparameters were optimized using grid search and early stopping based on validation AUC. Class imbalance was addressed using focal loss for neural models and class weighted training for classical ones.

# 4. Experiments

This section presents the experimental design used to evaluate the proposed models, including dataset characteristics, evaluation metrics, and performance results.

## 4.1 Datasets

We used two datasets for benchmarking:

LendingClub (Public): A real world dataset containing information on over 900,000 loans issued between 2007 to 2018. It includes borrower demographics, loan characteristics, and monthly repayment records.

Proprietary Microfinance Dataset: Provided by a lending institution operating in West Africa, this dataset includes 30,000 borrowers with monthly transactional and repayment histories over 24 months.

Each dataset was split into training (70%), validation (15%), and test (15%) sets, ensuring temporal consistency (training on earlier loans, testing on later ones).

## 4.2 Evaluation Metrics

We used the following metrics:

- AUC-ROC: Measures discriminative ability across thresholds.
- Log-Loss: Evaluates probabilistic accuracy of predictions.
- F1-Score: Focuses on model precision and recall in the presence of class imbalance.
- Calibration Error: Assesses how well the predicted probabilities align with observed outcomes.

## 4.3 Results

| Model | AUC | F1-score | Log-loss | Calibration Error |
|---|---|---|---|---|
| Logistic Regression | 0.79 | 0.65 | 0.41 | 0.083 |
| XGBoost | 0.82 | 0.69 | 0.39 | 0.072 |
| GRU | 0.84 | 0.71 | 0.36 | 0.064 |
| LSTM | 0.84 | 0.72 | 0.35 | 0.060 |
| Transformer | 0.87 | 0.74 | 0.32 | 0.054 |
| Ensemble | 0.86 | 0.73 | 0.33 | 0.056 |

# 5. Discussion

The results confirm that deep learning architectures offer significant advantages over traditional models in capturing borrower risk. Among the models tested, Transformer based architectures consistently achieved the highest performance across all evaluation metrics. This can be attributed to their ability to model long term dependencies and global sequence relationships through self attention mechanisms.

Recurrent models like LSTM and GRU also performed well, especially when sequences were short to medium in length. However, they showed limitations in terms of computational cost and vanishing gradients when dealing with longer sequences.

Ensemble models provided a strong middle ground. By integrating the probabilistic outputs of neural models with the decision boundaries of logistic regression or tree based classifiers, the ensemble methods delivered high predictive performance while maintaining a degree of transparency and interpretability necessary for deployment in regulated environments.

One of the most important insights from our experiments was the role of calibration. While neural models offered high accuracy, their raw probabilities were often misaligned with real world default rates.

Ensembles and post-hoc calibration techniques helped mitigate this issue, making the models more actionable.

Another noteworthy finding was the robustness of Transformer models in the presence of noisy or incomplete data. Their ability to assign attention selectively allows them to “ignore” irrelevant or missing inputs more effectively than RNNs.

Qualitative feedback from financial analysts and credit officers involved in pilot deployments suggested that model explanations and feature importance visualizations (e.g., via SHAP) played a critical role in gaining institutional trust and buy-in.

# 6. Conclusion

This study demonstrates that advanced deep learning architectures specifically Transformers, LSTMs, and GRUs offer substantial improvements in credit risk assessment, especially when trained on sequential borrower data. Our comparative evaluation across public and proprietary datasets highlights that Transformers outperform traditional and recurrent models in predictive accuracy, calibration, and robustness.

The inclusion of ensemble strategies further enhances the usability of these models in real world applications, bridging the gap between performance and interpretability. Our work also underscores the importance of fairness, transparency, and operational considerations when deploying machine learning models in financial services.

In an era where lending decisions are becoming increasingly automated and data driven, leveraging deep learning for credit scoring presents a transformative opportunity. However, success will depend not only on technical performance but also on ethical deployment, continuous monitoring, and regulatory alignment.

Future research should explore self supervised pretraining for financial sequences, federated learning for data privacy, and explainable AI techniques tailored to credit scoring contexts. Ultimately, our findings lay the foundation for next-generation credit risk systems that are more inclusive, accurate, and aligned with societal values.

# 7. Fairness, Interpretability, and Ethical Considerations

As machine learning models play an increasing role in credit decision making, ensuring their fairness, interpretability, and ethical deployment becomes a critical imperative. Credit decisions have profound social and financial consequences, and models must avoid perpetuating historical biases or making decisions that cannot be justified or explained.

# 7.1 Fairness in Predictive Modeling

Credit data often contains embedded historical biases whether due to socio economic inequality, structural exclusion, or legacy decision making processes. If unchecked, machine learning models trained on such data can reinforce or even amplify these biases. For instance, a model might inadvertently associate zip codes, employment gaps, or proxy variables (like education level or job title) with higher risk, even if those factors correlate more strongly with systemic inequity than with individual default behavior.

To mitigate this, we employed fairness aware evaluation strategies. We examined model performance across sensitive demographic subgroups (e.g., gender, income brackets, geographic regions) and measured metrics such as demographic parity difference and equalized odds. We found that Transformer models, although more complex, exhibited less performance disparity across groups compared to XGBoost and logistic regression. We attribute this to their ability to focus on nuanced behavioral patterns rather than high level proxy features.

## 7.2 Interpretability and Explainability

Deep learning models are often criticized for their “black box” nature, especially in financial services where transparency is a regulatory requirement. To address this, we applied model explainability tools including SHAP (SHapley Additive exPlanations) and integrated gradients. These methods helped visualize the contribution of each feature (e.g., payment delay trend, monthly income volatility) to the final risk score. For instance, in one use case, the SHAP summary plots revealed that sudden increases in revolving balance usage and frequent late payments were the strongest drivers of elevated risk scores across the Transformer model. These insights were shared with loan officers, who found them aligned with intuition and policy guidelines.

## 7.3 Ethical and Legal Considerations

Beyond technical fairness and transparency, ethical deployment requires safeguards like human in the loop review systems, informed consent for data use, and periodic bias audits. We recommend institutions adopt model cards and decision documentation for every deployment, enabling regulatory auditability and internal accountability.

# 8. Robustness and Generalization Across Markets

A key challenge in credit risk modeling is that borrower behavior varies across countries, economic conditions, and loan products. A model trained on one market may fail catastrophically in another if not properly validated or adapted.

## 8.1 Cross-Market Generalization

To evaluate generalization, we trained models on LendingClub U.S. data and tested them on the proprietary microfinance dataset from West Africa. Traditional models like logistic regression and

XGBoost exhibited significant performance degradation (AUC dropped by 8–12 points), likely due to distributional shift in feature values and borrower profiles.

In contrast, Transformer models when fine tuned on a small sample of local data retained most of their predictive performance (AUC dropped by only ~ 4 points). This suggests that attention-based architectures are better suited for domain adaptation due to their flexible, non-linear modeling of interactions.

## 8.2 Stress Testing Under Economic Shocks

We also tested how models respond to macroeconomic shocks, simulating 20% unemployment, declining GDP, and credit tightening scenarios. We found that GRU and Transformer models were more robust to sudden data shifts, especially when trained with dropout and data augmentation techniques. This robustness makes them suitable for deployment in volatile or emerging markets.

# 9. Applications in Real-World Credit Systems

Translating academic models into operational credit systems involves more than just predictive accuracy it requires integration into workflows, trust from analysts, and compatibility with existing infrastructure.

## 9.1 Operational Deployment

We collaborated with a fintech lender to integrate Transformer based models into their batch scoring pipeline. The models were deployed using TensorFlow Serving and scheduled via Apache Airflow. Risk scores were fed into dashboards used by credit analysts, who then reviewed and adjusted decisions based on both scores and SHAP generated explanations.

The scoring system evaluated over 12,000 loan applications monthly, reducing manual review volume by 35% while improving turnaround time by 28%.

## 9.2 Portfolio Outcomes and Feedback

After 3 months of deployment, default rates among newly accepted borrowers dropped by 11% compared to pre deployment cohorts, without a decrease in loan approval rates. Analysts reported greater confidence in the automated system due to its transparency and consistency. The explainability features were particularly helpful in edge cases involving gig economy workers and variable income borrowers.

This real world validation demonstrates that deep learning models when carefully integrated and monitored can be trusted in high-stakes financial environments.

# 10. Limitations and Future Directions

Despite promising results, this research has several limitations that should be acknowledged.

## 10.1 Data and Resource Constraints

Deep learning models require large volumes of well labeled sequential data. Institutions with limited digital infrastructure or sparse borrower histories may struggle to extract full value from these methods. Additionally, training and tuning deep architectures require substantial compute resources, which may be impractical for smaller lenders.

## 10.2 Interpretability vs. Complexity

There remains a tradeoff between predictive performance and interpretability. Although we mitigated this using SHAP and ensembles, there is still a gap in full transparency particularly when legal justifications for credit denial are required.

## 10.3 Future Work

Future research should explore:

- Self-supervised pretraining on financial sequences to reduce label dependence.
- Federated learning to train models across institutions without sharing raw data.
- Causal modeling to identify drivers of creditworthiness beyond correlation.
- Fairness-constrained optimization, balancing accuracy and equity in deployment.

These directions aim to ensure that deep learning models are not only accurate but also responsible and inclusive.

# 11. Conclusion (Extended)

This study set out to evaluate the impact of modern deep learning architectures on credit risk assessment, with a focus on sequential borrower data. Through a comparative analysis of LSTM, GRU, Transformer, and ensemble models, we demonstrated that deep learning especially attention-based Transformers offers significant improvements over traditional scoring methods.

These models excel not just in accuracy, but also in robustness, generalization, and dynamic learning from behavioral trends. Moreover, when combined with explainability tools and fairness audits, they become viable candidates for real world use in responsible lending ecosystems.

Beyond model performance, this research underscores a broader theme: the future of credit scoring lies in models that are adaptive, fair, and human-aligned. As financial services become increasingly data rich, it is crucial that the models we deploy are not only intelligent but also transparent and just.

Our work contributes both methodological insights and practical pathways for integrating AI into mission critical financial decision-making. In doing so, we hope to support a global movement toward more

inclusive, data driven, and accountable financial systems.

## References


1. Hand DJ, Henley WE (1997) Statistical classification methods in consumer credit scoring: A review. J Royal Stat Society: Ser A 160(3):523–541
2. Lessmann S, Baesens B, Seow HV, Thomas LC (2015) Benchmarking state-of-the-art classification algorithms for credit scoring. Eur J Oper Res 247(1):124–136
3. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD Conference* (pp. 785–794)
4. Jurgovsky J, Granitzer M, Ziegler K et al (2018) Sequence classification for credit-card fraud detection. Expert Syst Appl 100:234–245
5. Fischer T, Krauss C (2018) Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res 270(2):654–669
6. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008
7. Huang X, Khetan A, Cvitkovic M, Karnin Z (2020) TabTransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*
8. Zhang D, Zhou L, Zhai X (2020) Explainable AI for credit scoring: Balancing accuracy and interpretability. Financial Innov 6(1):25
9. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774
10. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Adv Neural Inf Process Syst 29:3315–3323
11. Galindo J, Tamayo P (2000) Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. Comput Econ 15(1):107–143
12. Baesens B, Setiono R, Mues C, Vanthienen J (2003) Benchmarking classification models for credit scoring. J Oper Res Soc 54(6):627–635
13. Ariza-Garzón MJ, Arroyo J, Caparrini G (2022) Credit scoring using attention-based deep learning models. Appl Soft Comput 113:107926
14. Khandani AE, Kim AJ, Lo AW (2010) Consumer credit-risk models via machine-learning algorithms. J Banking Finance 34(11):2767–2787
15. Sobhani S, Duman E (2021) Deep learning for credit scoring: Do attention mechanisms improve model interpretability? J Artif Intell Res 71:487–515
16. Louzada F, Ara A, Fernandes GB (2016) Classification methods applied to credit scoring: Systematic review and overall comparison. Surv Oper Res Manage Sci 21(2):117–134
17. Bastani H, Ascarza E, Choudhury P (2019) Predicting customer behavior using deep learning: The role of sequence modeling in credit and churn prediction. *Harvard Business School Working Paper*

No. 20–034

18. Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference* (pp. 1135–1144)
19. Bruckner T, Lehmann S, Urban J (2023) A Transformer-based architecture for default prediction: Empirical evidence from European lending markets. J Risk Financial Manage 16(2):85
20. Dastile XN, Celik T, Potsane M (2020) Statistical and machine learning models in credit scoring: A systematic literature survey. Appl Soft Comput 91:106263