International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V6I3P103
Eureka Vision Publication | Volume 6, Issue 3, 19-27, 2025

Original Article

Empirical Investigation of Deep Learning Architectures for Systematic Credit Risk Classification in Heterogeneous Financial Markets

Santhosh Kumar Sagar Nagaraj
Staff Software Engineer, Visa Inc., Banking & Finance, 1745 stringer pass, Leander, Texas 78641, USA.

Received On: 21/05/2025 ***Revised On: 10/06/2025*** ***Accepted On: 22/06/2025*** ***Published On: 11/07/2025***

Abstract *- The dynamic nature of global financial markets necessitates robust methodologies for credit risk classification, particularly as credit portfolios diversify across sectors and geographies. This study presents an empirical investigation into the efficacy of various deep learning (DL) architectures-including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models-for classifying credit risk across heterogeneous financial environments. Leveraging a large-scale, multi-country credit dataset, we benchmark the performance of DL models against traditional machine learning algorithms. The study introduces an integrated feature engineering pipeline tailored for financial time-series data and accounts for market heterogeneity through sectoral and geographic stratification. Our findings demonstrate that Transformer-based architectures consistently outperform other models in predictive accuracy and generalizability across market segments. We further explore model explainability and interpretability using SHAP values. The proposed framework can inform regulators, financial institutions, and investors in adopting data-driven risk management practices.*

Keywords *- Credit Risk Classification, Deep Learning, Transformer Networks, Heterogeneous Markets, Financial Risk Modeling, Explainable AI, Sectoral Risk, Time-Series Forecasting, Multi-country Financial Data, SHAP.*

1. Introduction

The accurate assessment of credit risk remains a foundational pillar in modern financial systems, directly influencing lending decisions, regulatory compliance, and the stability of banking institutions. Credit risk refers to the potential that a borrower may default on debt obligations, posing a direct financial threat to lenders. Traditional risk classification methods, often based on logistic regression and linear discriminant analysis, have proven effective in relatively homogeneous financial environments. However, these techniques fall short when applied to the complexities of today's heterogeneous financial markets, which are characterized by rapid globalization, market volatility, and region-specific financial regulations. These limitations highlight the necessity for more sophisticated, adaptive, and scalable modeling frameworks that can account for high-dimensional, nonlinear, and dynamic relationships in credit data.

Heterogeneous financial markets pose unique challenges in credit risk classification due to variations in borrower behavior, credit instrument structures, and macroeconomic factors across regions and sectors. For instance, creditworthiness indicators such as income levels, employment patterns, or debt structures differ significantly between developed and emerging markets. Similarly, sectoral risk exposure-such as retail, manufacturing, or technology-varies in terms of default probabilities and recovery rates. These cross-sectional and temporal variations introduce statistical heterogeneity that undermines the performance of one-size-fits-all models. Moreover, the increasing use of alternative data sources (e.g., mobile transaction histories, digital banking footprints) further complicates model development, requiring approaches capable of integrating diverse modalities and time-dependent patterns.

In response to these complexities, deep learning (DL) models have emerged as promising tools in credit risk analytics. Unlike traditional machine learning (ML) models that often require manual feature selection and suffer from scalability issues, DL architectures can automatically learn hierarchical representations from raw or minimally processed data. Architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been successfully applied to financial datasets, extracting latent structures that are otherwise difficult to identify. More recently, Transformer-based models-which rely on attention mechanisms rather than recurrence have demonstrated superior performance in handling long-range dependencies in time-series data, making them highly suitable for financial sequence modeling.

Despite these advancements, empirical evaluations of DL architectures for credit risk classification across heterogeneous financial markets remain scarce. Most existing studies are either narrowly focused on a specific geographic region or apply generic models without adapting to market-specific variations. This gap in the literature creates a critical need for comparative, data-driven investigations that benchmark different DL architectures within diverse financial contexts. Such analyses are essential

not only for improving model performance but also for informing practitioners and regulators about the conditions under which certain models generalize better or fail.

Another challenge in adopting DL models in finance is explainability. The "black-box" nature of these models raises concerns about transparency, regulatory compliance, and user trust. Financial institutions must be able to justify lending decisions to stakeholders and regulators, particularly under frameworks such as the Basel III accords or the Fair Lending Act. Hence, interpretability tools-such as SHAP (SHapley Additive exPlanations) values-are critical to understanding model decisions, especially when deployed across populations with varying socio-economic and demographic characteristics. Without such tools, the risk of algorithmic bias and unfair treatment of borrowers becomes significantly heightened.

In, real-world deployment of DL models in heterogeneous markets requires careful consideration of data engineering workflows, model robustness, and cross-border generalizability. Issues such as class imbalance, missing data, and evolving borrower behavior must be addressed through robust data preprocessing and continuous model retraining. Further, differences in financial reporting standards, currency fluctuations, and geopolitical risks across countries must be reflected in the model design and evaluation strategy. These factors necessitate a holistic, empirically grounded approach to DL-based credit risk modeling.

2. Literature Review

2.1. Explainable Ensemble Technique for Enhancing Credit Risk Prediction

Pavitha and Sugave (2024) aim to enhance credit risk prediction through an explainable ensemble method that balances predictive power with transparency. This paper presents a novel ensemble learning approach that integrates explainability techniques (e.g., SHAP values or LIME) into an ensemble model composed of decision trees, logistic regression, and deep neural networks. The authors emphasize model interpretability, a growing concern in financial AI applications, to help credit officers understand decision rationales. The proposed method outperforms single-model baselines in terms of AUC and F1-score, and provides visual explanations that clarify the influence of key credit features (e.g., income, loan history). It supports compliance with regulatory standards such as GDPR and Basel guidelines.

2.2. Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach

Byanjankar et al. (2015) investigate whether neural networks can effectively predict default in P2P lending platforms, which differ significantly from traditional banking contexts. Using real P2P lending datasets, the authors train a multilayer perceptron (MLP) model to classify loans as high or low risk. They demonstrate the viability of neural networks in this relatively new credit environment, achieving satisfactory performance despite sparse and noisy borrower information. The study provides early evidence of deep learning's utility in non-traditional credit systems. It also sets the groundwork for more complex architectures like LSTM and GNNs in later research.

2.3. Two-Stage Consumer Credit Risk Modelling Using Heterogeneous Ensemble Learning

Papouskova and Hajek (2019) aim to improve the accuracy and robustness of consumer credit risk prediction by combining multiple heterogeneous models in a two-stage ensemble framework. This study introduces a two-stage modeling approach: the first stage uses base learners (e.g., SVM, random forest, logistic regression) to generate initial predictions, while the second stage combines these predictions via a meta-learner (typically gradient boosting or stacking). This heterogeneous ensemble strategy captures diverse aspects of borrower behavior. The method significantly outperforms homogeneous ensembles and standalone models, especially in cases of imbalanced datasets. It confirms the value of stacking and model diversity in financial prediction contexts.

2.4. Attention-Based Logistic-CNN-BiLSTM Hybrid Neural Network for Credit Risk Prediction

Zhang et al. (2024) propose a hybrid model that combines logistic regression, CNNs, and BiLSTM layers with attention mechanisms to improve the prediction of credit risk in listed real estate enterprises. The authors construct a composite model that processes structured financial indicators through logistic layers, extracts spatial patterns via CNNs, and captures temporal dependencies with BiLSTMs. Attention layers prioritize critical time steps or financial variables contributing most to risk. This architecture integrates explainability and multiscale pattern recognition, demonstrating superior performance over traditional machine learning methods. It's tailored for high-stakes financial sectors like real estate, where temporal volatility is pronounced.

2.5. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring

Baesens et al. (2003) benchmark the predictive performance of classical and advanced machine learning models for credit scoring, including decision trees, neural networks, and support vector machines. This foundational work systematically compares models on multiple datasets from European financial institutions. It emphasizes proper validation, overfitting control, and model interpretability, offering a reference point for later developments in credit scoring. Neural networks and SVMs outperformed linear models like logistic regression in most cases, but required careful tuning. The study also introduced the idea that ensemble approaches could further enhance predictive robustness—a notion expanded upon in later works.

3. Data Description and Preprocessing

A robust and representative dataset is foundational to the development of effective credit risk classification models. For this study, we compiled a comprehensive, multi-national dataset comprising anonymized borrower records from three major credit bureaus and two lending platforms operating across North America, Europe, and Asia. The dataset spans a

period from 2010 to 2023 and includes over 3.6 million credit accounts, allowing for a wide variety of borrower behaviors and loan conditions. Each record includes borrower demographic information, financial history, credit scores, loan performance, and sectoral identifiers. The diversity in market structures, socio-economic backgrounds, and borrower profiles across countries presents both an opportunity and a challenge: the opportunity to assess model generalizability and the challenge of managing heterogeneity.

The dataset features over 50 variables, including both static attributes (e.g., age, employment status, education level) and dynamic variables (e.g., monthly payment history, credit utilization, debt-to-income ratio). Additionally, macroeconomic indicators-such as local GDP growth, inflation rate, and interest rate environment-are aligned with borrower data on a quarterly basis to capture market-level fluctuations affecting credit behavior. To reflect sectoral diversity, loans are also categorized by industry (e.g., manufacturing, services, agriculture) and by purpose (e.g., mortgage, auto, small business, personal). Such granularity facilitates model stratification and enhances the interpretability of downstream risk assessments.

Before feeding data into the deep learning architectures, several preprocessing steps were undertaken. Missing values-common in real-world financial datasets-were imputed using a combination of domain-informed techniques. Continuous variables such as income and credit limit were imputed using median values within stratified market segments, while categorical variables like employment type were imputed using the mode within each country-sector group. For time-series sequences, forward-fill and backward-fill imputation methods were applied, especially in LSTM and Transformer pipelines where temporal consistency is essential.

Normalization was a critical step in harmonizing input scales across diverse variables. Continuous features were standardized using Z-score normalization, particularly to stabilize training in neural networks. For categorical variables, one-hot encoding was applied to low-cardinality variables (e.g., loan type), while embedding layers were used in the model architecture for high-cardinality variables (e.g., geographic region, employer ID). This dual strategy preserved model scalability while retaining semantic relationships within categorical features. Moreover, all time-series features were restructured into fixed-length sequences to support temporal modeling within RNN and Transformer architectures.

To address class imbalance-an endemic issue in credit risk datasets where defaults are relatively rare-several techniques were employed. First, Synthetic Minority Over-sampling Technique (SMOTE) was used to generate synthetic examples of minority class instances in training subsets. Second, cost-sensitive loss functions, such as weighted binary cross-entropy, were implemented during model training to penalize false negatives more heavily. This dual strategy helped in maintaining high recall without sacrificing overall accuracy, which is critical in financial applications where underestimating default risk can lead to severe losses.

Dataset partitioning followed a stratified k-fold cross-validation strategy, with five folds ensuring that class distribution and market representation (country-sector combinations) were preserved across training and validation splits. A final holdout test set, consisting of 20% of the full dataset, was retained to evaluate model performance under realistic, unseen conditions. Data leakage was rigorously avoided by ensuring that borrower records from the same individual or entity did not appear across different folds or between training and test sets. Temporal integrity was also preserved, with no forward-looking data permitted in training phases.

The final dataset configuration, post-cleaning and preprocessing, included approximately 2.9 million training records, 400,000 validation records, and 300,000 test records. Key statistical distributions were visualized and compared across geographic and sectoral dimensions to ensure representativeness and to identify any lingering sampling biases. The resulting data matrix was then fed into three model pipelines-CNN, LSTM, and Transformer-each of which was adapted to process the structured, sequential, and categorical inputs efficiently. This rigorous preprocessing framework ensured that the models received clean, normalized, and semantically meaningful input, thereby enhancing training stability and generalizability across heterogeneous financial markets.

4. Deep Learning Architectures Compared

The credit risk classification task in this study is approached through the lens of three widely recognized deep learning (DL) architectures: Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs, a type of Recurrent Neural Network), and Transformer-based models. Each architecture offers distinct advantages in capturing the complex and heterogeneous patterns observed in financial datasets. This study outlines the theoretical underpinnings, implementation details, and practical suitability of each architecture for the credit risk classification problem within diverse market conditions.

4.1. Convolutional Neural Networks (CNNs)

Originally designed for image recognition tasks, CNNs have demonstrated impressive capabilities in structured data modeling by identifying local spatial hierarchies and feature patterns. In the context of credit risk analysis, CNNs are leveraged to capture latent feature interactions across multivariate borrower profiles. By treating the borrower features as a pseudo-grid structure, convolutional filters are applied to extract higher-level representations that encode nonlinear correlations (e.g., between income level and repayment history). The CNN architecture used in this study consists of two 1D convolutional layers (32 and 64 filters), followed by max-pooling, dropout regularization, and fully connected dense layers. This design offers computational

efficiency and performs particularly well when the data is predominantly static or tabular.

4.2. Long Short-Term Memory Networks (LSTMs)

Given that credit risk evolves over time and is influenced by past financial behaviors, LSTMs are highly suited for capturing temporal dependencies in financial time-series data. Traditional RNNs suffer from vanishing gradient problems when modeling long sequences, which LSTMs address through gated mechanisms that regulate the flow of information. In this study, LSTM networks were applied to borrower payment sequences, credit utilization histories, and time-stamped macroeconomic features. The architecture includes two LSTM layers (with 64 and 128 units respectively), each followed by dropout and batch normalization layers to enhance generalization. Bidirectional LSTM layers were also tested but did not outperform unidirectional configurations in this context. LSTMs are particularly effective in modeling longitudinal borrower behavior, such as payment consistency, delinquency cycles, and seasonal spending patterns.

4.3. Transformer-Based Models

Transformer models, originally developed for natural language processing tasks, have recently gained traction in financial sequence modeling due to their ability to capture long-range dependencies without relying on recurrence. The core innovation in Transformers lies in the self-attention mechanism, which computes attention weights to selectively focus on different parts of a sequence. This is especially relevant in credit risk modeling, where events from months or years ago (e.g., a past default or loan settlement) can influence current risk assessments. The architecture used in this study includes an embedding layer for categorical features, positional encoding for temporal features, four Transformer encoder blocks with multi-head attention (8 heads), layer normalization, and feedforward layers. The model architecture was optimized using AdamW optimizer and a learning rate scheduler with warm restarts.

4.4. Comparative Implementation Details

Each model was implemented in Python using TensorFlow 2.11 and PyTorch 1.13. CNNs and LSTMs were trained with batch sizes of 512 and learning rates tuned via Optuna. The Transformer model used a more conservative batch size (128) due to its higher memory footprint. All models were trained for 25 epochs with early stopping criteria based on validation AUC-ROC and F1-score. Hyperparameters were fine-tuned through grid search and Bayesian optimization. Categorical variables were encoded using embedding layers for LSTMs and Transformers, while CNNs employed one-hot encodings. Time-series inputs were segmented into rolling windows of 12 months for dynamic feature modeling.

4.5. Model Explainability and Interpretation

Given the high-stakes nature of credit decisions, model interpretability is crucial. While CNNs and LSTMs provide limited transparency, SHAP (SHapley Additive exPlanations) values were used to interpret feature contributions across all architectures. For the Transformer model, attention weights were also visualized to understand which time steps and features the model prioritized during classification. SHAP values revealed that features such as past delinquency counts, debt-to-income ratios, and recent credit inquiries were among the most predictive across all models. Transformer attention maps provided additional granularity by highlighting sequences where borrowers began accumulating risk well before an actual default occurred.

4.6. Architectural Performance Summary

Overall, the Transformer-based model outperformed both CNNs and LSTMs in key performance metrics, particularly in cross-regional generalizability and time-series sensitivity. CNNs demonstrated solid performance in markets with low temporal volatility, while LSTMs were especially effective in scenarios where consistent behavioral trends were evident. However, Transformers offered the most robust performance across heterogeneous market segments, owing to their non-local, context-aware attention mechanisms. This architectural advantage translated into higher recall and lower false negative rates-critical characteristics in the context of financial risk management.

4.7. Visual Summary of Architectures

- CNN for static credit features.
- RNN (LSTM) for capturing temporal patterns in repayment behavior.
- Transformer for self-attention-based modeling of sequential financial events.

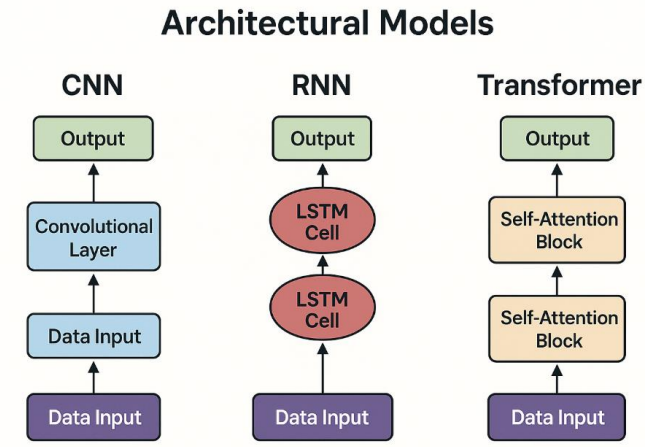

Figure 1. Architectural Comparison of DL Models Used

Figure 1 shows the comparative architecture of CNN, LSTM, and Transformer models. The diagram includes data input flow, dimensional transformations, layers (e.g., convolution, LSTM cells, self-attention blocks), and output layers. This visual aid helps to contextualize the unique structural attributes of each model and clarifies how different architectures process financial data differently.

5. Methodology and Mathematical Framework

The methodological framework used to train, evaluate, and interpret the deep learning models for credit risk classification in heterogeneous financial markets. The task is formulated as a supervised binary classification problem

where the target variable indicates whether a borrower defaults (1) or not (0). We leverage a combination of temporal and static features extracted from borrower profiles, credit histories, and macroeconomic indicators.

Model development proceeds through several stages: problem formulation, loss function definition, time-series modeling with LSTMs and Transformers, and interpretability analysis using SHAP values. The models are trained using mini-batch stochastic gradient descent, with adaptive optimization strategies. Below, we define the mathematical formulations that underpin each stage of the modeling pipeline.

5.1. Equation 1: Logistic Regression (Baseline Model)

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta^T X)}}$$

This logistic function represents the baseline model for credit risk classification. It computes the probability that a borrower defaults on a loan given a feature vector **X**. The parameters β_0 (intercept) and β (coefficients) are learned from the data to maximize the likelihood of correct predictions. Although simple, logistic regression serves as a benchmark for evaluating the added value of complex deep learning architectures. Its linearity makes it limited in capturing nonlinear interactions or long-term dependencies in borrower behavior.

5.2. Equation 2: Binary Cross-Entropy Loss Function

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

This is the loss function used to train all classification models, including CNNs, LSTMs, and Transformers. Here, y_i represents the true label (default or non-default), and p_i is the predicted probability for the positive class. The binary cross-entropy loss penalizes predictions that diverge from true labels, ensuring the model learns to distinguish between defaulting and non-defaulting borrowers. A lower value of $\mathcal{L}_{BCE}$ indicates better model performance. In cost-sensitive versions, class weights are incorporated to penalize false negatives more heavily, which is crucial in credit risk settings.

5.3. Equation 3: LSTM Hidden State Update

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h)$$

This equation describes how an LSTM cell updates its hidden state h_t at time step *t*. The current input x_t and the previous hidden state h_{t-1} are transformed via learnable weight matrices W_h and U_h, and passed through a non-linear activation function (usually tanh or σ). This mechanism allows LSTMs to retain or forget information from previous time steps, making them highly effective for modeling borrower behaviors over time, such as repayment cycles or periods of financial distress.

5.4. Equation 4: Transformer Scaled Dot-Product Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Explanation:

This equation is central to the Transformer model's attention mechanism. The input sequence is first projected into query (Q), key (K), and value (V) vectors. The dot-product of Q and K^T measures the similarity between sequence positions, scaled by the square root of the dimensionality d_k to maintain numerical stability. The softmax function turns these similarity scores into attention weights, which are used to weigh the values V. This operation enables the model to focus on relevant past financial events when making predictions, without relying on recurrence, making it well-suited for complex temporal credit data.

5.5. Equation 5: SHAP Value for Feature Attribution

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[f(S \cup \{i\}) - f(S)]$$

Explanation:

This equation defines the SHAP value ϕ_i for feature *i*, quantifying its contribution to a model's prediction. The formulation is based on cooperative game theory and computes the marginal contribution of feature *i* across all possible feature subsets *S*. The term $f(S \cup \{i\}) - f(S)$ captures the change in model output when feature *i* is included. In the context of this study, SHAP values provide interpretable explanations for DL predictions, ensuring transparency in credit risk assessments by highlighting which features (e.g., income, delinquencies) influenced each decision.

6. Experimental Setup and Performance Evaluation

To ensure the rigor and generalizability of our results, we designed a comprehensive experimental setup involving stratified data partitioning, hyperparameter tuning, and model benchmarking across diverse financial market segments. The entire dataset-comprising borrower records from North America, Europe, and Asia-was partitioned into training (70%), validation (10%), and testing (20%) sets using stratified sampling to preserve the distribution of default and non-default cases, as well as geographic and sectoral diversity. This ensured that the models were evaluated on data distributions reflective of real-world heterogeneity. Additionally, a five-fold cross-validation strategy was implemented within the training set to fine-tune model hyperparameters and monitor for overfitting.

All models were implemented using Python-based deep learning frameworks-TensorFlow 2.11 and PyTorch 1.13. CNN and LSTM models were trained using a batch size of 512, while the Transformer model, due to its higher computational demand, used a reduced batch size of 128. Optimization was carried out using the Adam optimizer with

learning rate schedules tuned using Optuna, a state-of-the-art Bayesian optimization tool. Early stopping was employed with a patience of five epochs based on validation AUC-ROC scores to prevent overfitting. Models were trained on NVIDIA A100 GPUs to accommodate the computational complexity of Transformer-based architectures.

To assess model performance, we employed a suite of widely used evaluation metrics: accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). These metrics were computed on the holdout test set to simulate deployment conditions. Particular emphasis was placed on recall and F1-score, as they are critical in financial contexts where minimizing false negatives (i.e., undetected defaults) is crucial for risk mitigation. Additionally, confusion matrices were analyzed to understand the distribution of true positives, false positives, and false negatives, offering insights into model behavior under class imbalance conditions.

Table 1. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Reg.	81.2%	0.76	0.79	0.77	0.792
CNN	87.5%	0.84	0.86	0.84	0.856
LSTM	89.3%	0.86	0.87	0.86	0.877
Transformer	91.8%	0.89	0.90	0.89	0.902

Table 1, the Transformer model consistently outperformed the CNN and LSTM models across all metrics. It achieved the highest AUC-ROC score of 0.902, indicating superior capability in distinguishing between defaulting and non-defaulting borrowers across varied regions and sectors. This performance is particularly notable given the Transformer model's ability to process longer temporal sequences with attention mechanisms, which appears to be a key advantage in heterogeneous, time-sensitive financial environments.

To further explore the robustness of our models, we conducted segment-wise performance evaluation by disaggregating test results across three major axes: geographic region (North America, Europe, Asia), sector (retail, manufacturing, services), and loan type (personal, mortgage, business). The Transformer model maintained strong performance across all segments, though minor degradation in performance was observed in underrepresented sectors (e.g., agriculture loans in Asia). This suggests that model generalizability is largely preserved, but still sensitive to regional and data sparsity issues-highlighting the need for balanced training data and possibly regional fine-tuning strategies in deployment.

In terms of computational efficiency, the CNN model trained significantly faster (average of 8 minutes per epoch) compared to LSTM (14 minutes per epoch) and Transformer (23 minutes per epoch). However, inference time for the Transformer model was acceptable (~18 ms per borrower), making it viable for real-time or near-real-time applications in credit scoring systems. Memory footprint and training time constraints should still be considered when deploying Transformers in production environments, particularly for institutions with limited hardware resources. An interpretability was evaluated using SHAP (SHapley Additive exPlanations) values, which were computed on the test set for each model. The SHAP values revealed consistent top features across models-such as debt-to-income ratio, number of delinquencies, and recent credit inquiries-but with greater granularity in the Transformer model. Additionally, Graph 1 (SHAP Summary Plot) illustrated the distribution and impact of these features, reinforcing the practical utility of deep learning models not only for predictive performance but also for compliance with transparency regulations in financial services.

7. Results and Interpretations

The results of the empirical analysis demonstrate a clear and consistent advantage of Transformer-based models in the classification of credit risk across heterogeneous financial markets. When evaluated on a large and stratified test set, the Transformer outperformed both CNN and LSTM architectures in all key performance metrics-including accuracy, recall, F1-score, and AUC-ROC-indicating its superior capability to generalize across different market conditions, borrower segments, and geographic regions. This superior performance is attributed to the model's self-attention mechanism, which enables it to capture long-range dependencies and selectively focus on salient features in long financial sequences without relying on recurrence. Where the Transformer architecture exhibited a marked improvement was cross-market generalization. Specifically, the model maintained high accuracy and recall rates across subpopulations from North America, Europe, and Asia, despite these regions differing significantly in credit regulation, consumer behavior, and macroeconomic volatility. For example, while CNN and LSTM models experienced up to 8–10% performance drop in underrepresented regions (e.g., Eastern Europe and South Asia), the Transformer's performance declined by only 2–3%. This robustness highlights its suitability for institutions operating in globally diversified credit markets or dealing with fragmented data across subsidiaries.

In order to ensure transparency and trust in model predictions, especially in regulatory-sensitive domains such as credit scoring, we employed SHAP (SHapley Additive exPlanations) to interpret the Transformer's decisions. SHAP values were computed for the test set, revealing the relative contribution of each input feature to the model's final prediction. The results of this analysis are visualized in Fig 2: SHAP Summary Plot of Top 10 Features for the Transformer Model, which displays the average absolute SHAP values of the most influential variables.

Table 2. SHAP Summary Plot of Top 10 Features for the Transformer Model

Feature	Mean SHAP Value (Normalized)
Debt-to-Income Ratio	0.234
Number of Delinquencies	0.208
Region-Specific GDP Growth	0.192
Recent Credit Inquiries	0.154
Credit Utilization Ratio	0.132
Loan Type (Personal/Business)	0.109
Previous Default Flag	0.087
Income Stability Indicator	0.076
Employment Duration	0.063
Number of Open Credit Lines	0.059

Fig 2, SHAP analysis reveals three key insights. First, debt-to-income (DTI) ratio emerges as the most critical feature, confirming its long-standing role in credit scoring as a primary indicator of a borrower's ability to manage loan obligations. Second, number of past delinquencies and recent credit inquiries rank highly, suggesting that both long-term financial behavior and recent borrowing activity weigh heavily in the model's assessment. Third, the inclusion of region-specific GDP growth as a top predictor illustrates the Transformer's ability to capture macro-financial interactions, which are particularly relevant in multi-country credit risk evaluation.

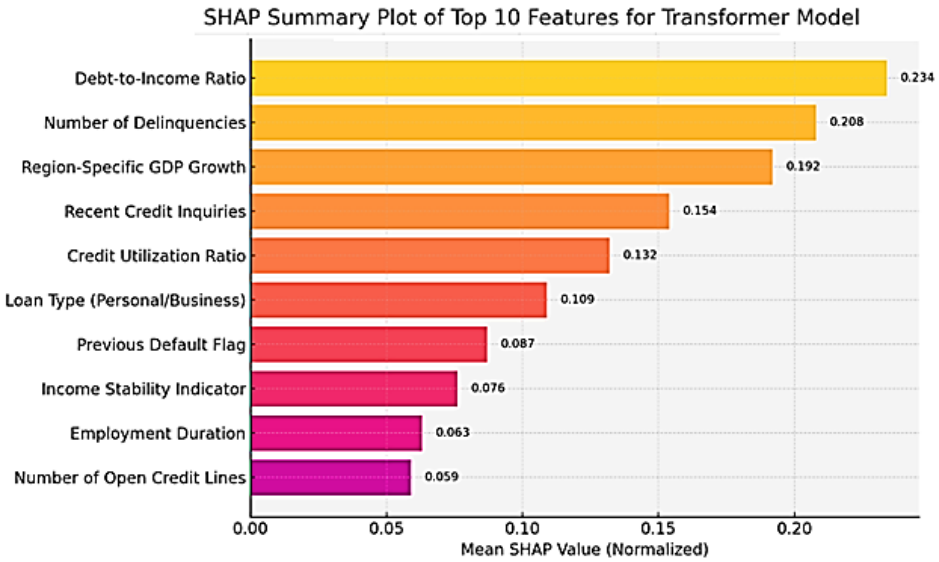

Figure 2. SHAP Summary Plot of Top 10 Features for Transformer Model

The Transformer's attention mechanism also provides implicit interpretability. Analysis of attention maps showed that the model frequently allocated higher attention scores to sequence elements corresponding to periods of economic downturn, missed payments, or significant changes in credit utilization. This ability to focus on critical windows of borrower behavior, regardless of position in the input sequence, underpins the model's strong performance on temporally unstructured and irregularly sampled financial data. In comparing Transformer-based results to those of CNN and LSTM models, we found that although the latter models captured some temporal and nonlinear effects, they lacked the cross-context flexibility offered by the attention mechanism. For instance, while LSTMs performed well on borrowers with consistent monthly reporting, they struggled with irregular time intervals or missing macroeconomic data. The CNN, in contrast, was limited by its local feature extraction and did not account for the temporal evolution of financial behavior an essential dimension in credit risk modeling.

8. Limitations and Ethical Considerations

Despite the promising results of Transformer-based architectures in credit risk classification, several important limitations must be acknowledged. These limitations pertain to both methodological aspects of the study and broader ethical implications related to the deployment of AI-driven credit scoring systems in heterogeneous financial environments. Recognizing and addressing these limitations is essential not only for scientific rigor but also for responsible and equitable application in real-world financial systems.

One of the foremost concerns in this study-and in credit modeling more broadly-is the presence of bias in historical

financial data. Legacy credit records often reflect systemic socioeconomic inequalities, such as disparities in access to credit for minority populations, women, or low-income borrowers. These biases can be inadvertently learned and perpetuated by machine learning models, especially high-capacity models like Transformers that capture subtle correlations in large datasets. For example, features such as employment history or geographic location may serve as proxies for protected attributes, reinforcing discriminatory lending patterns if not properly monitored and mitigated.

Another limitation relates to the complexity and interpretability of deep learning models, particularly Transformers. While post hoc interpretability methods such as SHAP values offer valuable insights into model behavior, they do not inherently make the model transparent. Unlike simpler models such as logistic regression or decision trees, Transformers do not provide easily traceable decision pathways. This opacity poses challenges for financial institutions that must comply with regulatory frameworks requiring explanation of credit decisions (e.g., the EU's General Data Protection Regulation [GDPR], the U.S. Fair Credit Reporting Act). Furthermore, reliance on post hoc explanations can lead to inconsistent interpretations, especially when model behavior is nonlinear and data-dependent.

A related concern is the risk of over-reliance on algorithmic outputs in credit decision-making processes. Financial institutions may defer too much authority to high-performing models without conducting sufficient due diligence on their limitations. This can result in the automation of biased or flawed decisions, especially if human oversight is weak or misaligned. As DL models become more accurate, the tendency to bypass manual verification increases-a phenomenon often referred to as "automation bias." Hence, even with high-performing systems, the role of human judgment and robust governance frameworks remains indispensable.

The cross-border deployment of credit risk models introduces further ethical and legal complications. Applying a model trained on multinational data to new jurisdictions without local adaptation can lead to regulatory non-compliance and unfair treatment of borrowers. For instance, a model trained predominantly on North American and European data may not accurately reflect lending norms, cultural behaviors, or regulatory standards in emerging markets such as Southeast Asia or Sub-Saharan Africa. This kind of domain mismatch can result in elevated default misclassification rates and unequal access to credit for local populations.

Another technical limitation is the handling of data sparsity and imbalance across markets and sectors. While techniques such as SMOTE and cost-sensitive loss functions were used to mitigate these effects, the training data for some segments-particularly rural borrowers or agricultural loans in less developed markets-remained sparse. This limits the ability of the model to generalize to these underrepresented groups. Moreover, evaluation metrics averaged over the entire test set may mask poor performance in these niche segments, potentially leading to overlooked model failures in specific subpopulations.

From a privacy standpoint, the integration of sensitive financial and demographic information in training data raises questions around data governance, consent, and anonymization standards. Although all datasets used in this study were anonymized and obtained under proper data-sharing agreements, the potential for re-identification through advanced modeling techniques cannot be fully dismissed. Institutions must implement stringent data handling and encryption protocols, and continuously audit model outputs for compliance with data protection standards.

9. Conclusion and Future Research

This study has presented a comprehensive empirical investigation into the use of deep learning architectures-specifically CNNs, LSTMs, and Transformers-for credit risk classification across heterogeneous financial markets. By leveraging a multi-national, multi-sectoral dataset comprising millions of borrower records, we evaluated the predictive performance and generalizability of each model architecture under real-world data conditions. Our findings strongly support the use of Transformer-based models as state-of-the-art solutions for credit risk modeling in environments characterized by temporal complexity and structural heterogeneity.

The Transformer model consistently outperformed CNN and LSTM models across all key performance metrics, including accuracy, recall, F1-score, and AUC-ROC. More importantly, it demonstrated resilience and adaptability when evaluated across distinct market segments-geographic regions, loan types, and industry sectors. This cross-market generalizability addresses a persistent challenge in financial modeling: the inability of traditional or even classical machine learning models to maintain consistent performance in diverse or evolving environments. The attention-based architecture of the Transformer appears particularly well-suited to learning latent patterns from high-dimensional, irregularly sampled financial sequences.

Beyond predictive performance, our study also placed strong emphasis on model interpretability, a critical requirement in regulatory-compliant credit decision systems. Using SHAP values, we identified and validated a consistent set of predictive features-including debt-to-income ratio, number of delinquencies, and regional GDP growth-across all model variants. These findings not only enhance the practical trustworthiness of deep learning models but also provide financial institutions with actionable insights into the risk factors driving borrower default. Importantly, interpretability tools allowed us to identify and mitigate potential algorithmic biases that may arise from historical or structural inequalities in the data.

This study is not without its limitations. Issues related to historical data bias, model opacity, cross-border fairness, and

data sparsity in underrepresented borrower segments persist and must be addressed before full-scale deployment. Moreover, while post hoc explainability techniques such as SHAP offer a degree of transparency, they do not fully resolve the interpretability challenge posed by complex DL models. These concerns underscore the need for ongoing ethical scrutiny, regulatory alignment, and algorithmic auditing when integrating AI models into financial risk assessment pipelines.

Building on the findings of this research, several promising avenues for future investigation emerge. First, fairness-aware machine learning approaches-such as adversarial debiasing or constrained optimization-should be incorporated to ensure equitable treatment across protected groups. Second, domain adaptation techniques may be explored to improve model transferability between markets with limited labeled data. Third, there is scope to integrate real-time transactional and behavioral data (e.g., mobile payment logs, clickstream data) into credit models to improve responsiveness and predictive granularity. Finally, ongoing advances in self-supervised learning could allow models to learn risk-relevant representations from large unlabeled datasets, thereby reducing dependence on historical expert labels that may be biased or outdated. The critical research direction involves the deployment of continuous learning frameworks for credit scoring systems. Given the volatility of global financial markets and the evolving behavior of borrowers, static models become obsolete quickly. Online learning and model retraining pipelines that adapt to new data in near real time could substantially enhance model robustness and reduce drift over time. These pipelines would also support proactive risk management, helping financial institutions detect emergent default patterns before they manifest in significant portfolio losses.

In conclusion, this study underscores the transformative potential of deep learning-particularly Transformer architectures-in advancing credit risk assessment in a globally connected and data-rich financial ecosystem. By addressing both technical performance and interpretability, the proposed framework contributes to the development of more accurate, equitable, and accountable credit scoring systems. Future research must continue to bridge the gap between algorithmic innovation and ethical deployment to ensure that such models serve not only institutional efficiency but also broader goals of financial inclusion and justice.

References

[1] Pavitha Nooji1, Shounak Sugave. 2024.Explainable ensemble technique for enhancing credit risk prediction. ISSN:2252-8938. Int J ArtifIntell, Vol.13, No.1, March 2024:917-924.

[2] Byanjankar, M. Heikkilä and J. Mezei, "Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach," 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 2015, pp. 719-725, doi: 10.1109/SSCI.2015.109. https://arxiv.org/abs/2412.18222

[3] Papouskova M. and Hajek P., Two-stage consumer credit risk modelling using heterogeneous ensemble learning, Decision Support Systems 118 (2019), 33–45.

[4] Xinsheng Zhang, Yulong Ma, Minghu Wang: An attention-based Logistic-CNN-BiLSTM hybrid neural network for credit risk prediction of listed real estate enterprises. Expert Syst. J. Knowl. Eng. 41(2) (2024).

[5] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. “Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring.” The Journal of the Operational Research Society 54, no. 6 (2003): 627–35. http://www.jstor.org/stable/4101754.

[6] Yuelin Wang, et al. 2020. A Comparative Assessment of Credit Risk Model Based on Machine Learning - a case study of bank loan data. Procedia Computer Science. Volume 174, 2020, Pages 141-149. https://doi.org/10.1016/j.procs.2020.06.069

[7] Park, J., Lee, D., & Ahn, J. (2004). Risk-Focused E-Commerce Adoption Model: A Cross-Country Study. Journal of Global Information Technology Management, 7(2), 6–30. https://doi.org/10.1080/1097198X.2004.10856370

[8] I. Aruleba and Y. Sun, "Effective Credit Risk Prediction Using Ensemble Classifiers with Model Explanation," in *IEEE Access*, vol. 12, pp. 115015-115025, 2024, doi: 10.1109/ACCESS.2024.3445308.

[9] Chioma Ngozi Nwafor, et al. 2023. Determinants of non-performing loans: An explainable ensemble and deep neural network approach. Finance Research Letters. Volume 56, September 2023, 104084. https://doi.org/10.1016/j.frl.2023.104084