International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Generative AI and Dynamic Modeling for Real-Time Cloud-Based Credit Risk

Kandasamy Sellappan

Abstract: The high rate of the digital financial ecosystem development has revealed the structural weaknesses of the old credit risk systems based on the fixed models and batch-processing setting. These systems are unable to record the real-time borrower activities and are not flexible to macroeconomic fluctuations. In order to fill this gap in architecture, this research project suggests Generative-Dynamic Risk Architecture (GDRA). The aim of the research is to synthesize the progress in four existing siloed technological areas: cloud-native orchestration, Generative AI, transformer-based sequence modeling, and Explainable AI (XAI). Our comparative study provides a number of main conclusions. First, event-driven microservices can dramatically improve the resilience of the system by 40% of the Mean Time To Recovery (MTTR). Second, Generative Adversarial Networks (GANs) and diffusion models can effectively reduce extreme class imbalance and simulate stress situations and maintain data privacy. Third, the predictive core uses Transformer networks to learn long-range temporal dependencies, and it persistently achieves better results in default prediction than classic recurrent models. Last but not least, SHAP-based XAI integration will make these complex models adhere to the stringent regulatory transparency requirements. We find that the GDRA offers a needed paradigm shift that offers a constantly recalibrating, fault-tolerant infrastructure that balances high predictive accuracy with regulatory accountability in contemporary credit risk management.

Keywords: *Credit Risk Management, Synthetic Data Augmentation, Temporal Sequence Modeling, Event-Driven Architecture, Explainable AI (XAI)*

I. Introduction

The financial services industry is experiencing a structural change like never before due to the fast-growing real-time payment networks, mobile banking apps, peer-to-peer lending platforms, and embedded finance [25]. Within this fast, digital ecosystem, the conventional credit risk frameworks, which are mostly based on fixed Probability of Default (PD) and Loss Given Default (LGD) models, are structurally inappropriate to the huge and dynamic data landscapes they are trying to model [12]. The traditional financial institutions used to use traditional econometric techniques and monolithic frameworks that are clearly defined to work in batch-processing settings. Nonetheless, such legacy systems are characterized by serious latency in data, which does not allow recognizing revenue in real-time and does not support the ability to react timely to the situation. The complex, multi-dimensional, and time-varying behavioral patterns in the context of the contemporary flow of transactions are incredibly hard to model using the traditional rule-based heuristics and single-score models. As a result, solid dependence on the old batch-processed models in the current algorithmic lending environment inevitably results in a decline in scorecard quality, a slow reaction to risk, and an organizational failure to react proactively to macroeconomic fluctuations [12].

In order to get past the shortcomings of these legacy credit risk systems, recent literature has considered various advanced technological paradigms separately. To address the longstanding issue of extreme class imbalance in financial data, Generative Artificial Intelligence (GenAI) has been presented in the first place, namely Generative Adversarial Networks (GANs) and diffusion models. These generative models enable the institutions to generate realistic data of the minority classes' defaults and can simulate extreme

Independent Researcher, USA

macroeconomic stress conditions without violating data privacy [2]. Second, temporal sequence modeling with transformers has shown an enhanced ability to learn long-range dependence of behavior and irregular time periods in financial time series, greatly outperforming classic Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks in default prediction and fraud detection [9]. Third, microservices based on event-driven cloud-native architecture have become the foundation of resilient financial processing. These architectures can use asynchronous messaging, container orchestration (e.g., Kubernetes), and automated failover protocols to maintain high operational availability and quick recovery in case of system failures. Lastly, Explainable AI (XAI) integration, especially frameworks such as SHapley Additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), has been made an obligatory governance tool in order to unpack black-box machine learning predictions so that they could meet the requirements of strict regulatory transparency requirements, such as the European Union Artificial Intelligence Act [18].

Although there has been tremendous empirical progress in these personal areas, there remains a research gap that is critical. The problem that contemporary financial institutions face is not a technical one but an architectural one. A systematic review of the existing state of art shows that GenAI data augmentation, temporal sequence modeling, cloud-native orchestration, and algorithmic accountability are mostly discussed as discrete, siloed solutions. None of the reviewed frameworks have yet provided an effective way in which these separate pillars can be effectively combined into one, constantly-recalibrating credit risk architecture. Without such an integrated framework, financial entities are exposed to systemic risks, regulatory arbitrage, and infrastructural friction in an effort to roll out contemporary artificial intelligence on an enterprise level. A holistic system should also be capable of processing large volumes of data at high velocity, synthesizing infrequent risk events, in-flight updating sequential predictive models, and making the final credit decision fully auditable.

The paper discusses this critical gap in integration by proposing and systematically validating a synthesized architecture contribution, the Generative-Dynamic Risk Architecture (GDRA). The main goal of the study is to conceptualize an end-to-end, fault-tolerant credit risk framework that matches sophisticated generative and predictive capabilities to the real-time needs of the digital financial ecosystem. This study has triple explicit contributions. We then develop a layered GDRA architecture that structurally integrates a self-healing, event-driven data infrastructure with a generative synthetic oversampling engine and a dynamic transformer-based temporal modeling layer. Second, we create a continuous XAI interface within the temporal predictive pipeline to ensure algorithmic responsibility, reduce discriminatory bias, and meet regulatory transparency goals. Third, we give a comparative thematic synthesis of empirical benchmarks to confirm the theoretical performance, architectural resilience, and compliance alignment of the GDRA with legacy baselines [12].

The rest of this paper is organized in the following way. Section 2 presents a systematic literature review of the shortcomings of legacy credit systems and the siloed efforts in GenAI, temporal modeling, cloud-native architectures, and XAI and ends with a formal gap analysis. Section 3 describes the methodology, including the systematic search strategy, the criteria of thematic synthesis, and the design justification that the GDRA framework is based on. Section 4 contains the results that encompass the exposition of the structural layers of the GDRA and the performance analysis in the form of benchmarking of the predictive accuracy and resilience of the system. Section 5 talks about the implications of these findings for researchers, industry practitioners, and regulators, as well as the limitations and threats to the validity of the proposed architecture. Lastly, Section 6 provides a conclusion of the paper with an overview of fundamental contributions and future longitudinal research and harmonization of regulatory practices across jurisdictions.

II. Literature Review

The reorganization of credit risk management frameworks as the traditional and linear behavior of borrowers gives way to the high-velocity and digitalized financial ecosystem. The old systems, based heavily on the traditional models of Probability of Default (PD) and Loss Given Default (LGD), are structurally mismatched

with the real-time, multi-platform, and extremely volatile characteristics of contemporary financial settings. This literature review is a systematic review of literature on the technological pillars needed as next-generation risk systems: Generative AI to augment synthetic data, transformer-based temporal sequence prediction, cloud-native microservice orchestration, and Explainable AI (XAI) to comply with regulatory requirements. This section provides an overview of the individual achievements of these technologies and reveals the serious architectural gap that the proposed Generative-Dynamic Risk Architecture (GDRA) seeks to address.

A) Limitations of Legacy Credit Risk Systems

The traditional econometric and statistical models, including the Logistic Regression, have historically dominated credit risk scoring, as these are simpler and highly interpretable [19]. Nonetheless, these conventional mathematical and statistical models have serious constraints in handling large data volumes of high dimensions that define consumer behavior today. It has been shown empirically that the traditional models have a hard time representing complex non-linear relationships in financial data, which reduces their predictive power relative to more sophisticated Machine Learning (ML) strategies [18].

Moreover, historic enterprise applications in the financial industry have been based on monolithic systems and batch processing [12]. This batch-based paradigm is fundamentally inappropriate to current credit risk measurement because it creates a lot of latency in data, real-time recognition of revenues, and cannot easily make timely decisions in times of macroeconomic turmoil [12]. The monolithic systems also do not scale, are not responsive and flexible enough to deal with dynamic multi-channel transaction processing, and eventually cause scorecard degradation and model drift as borrower behaviors change more rapidly than the systems can re-optimize. Therefore, scholars suggest that it is time to abandon the concept of application scoring, and move to dynamic behavioral scoring that demands the use of architectures that can constantly process live streams of transactions [1].

B) Generative AI in Financial Data Contexts

The problem of extreme imbalance of classes is constantly encountered by researchers in the credit scoring and fraud detection fields; fraudulent transactions or loan defaults (the minority class) are infrequent compared to legal, performing accounts (the majority class) [20]. Classifiers trained on this type of imbalanced data tend to be strongly biased towards the majority class, strongly reducing false negatives [2]. To counter this, Generative AI, especially Generative Adversarial Networks (GANs), has become a better way of synthetic data oversampling [20].

GANs are based on a min-max game between two neural networks: a Generator (G) that takes random noise as input and generates data and a discriminator (D) that assesses the reality or fake nature of the sample [20]. Research shows that GANs have been effective at capturing the underlying joint distribution of highly complex financial data compared with traditional resampling algorithms, such as SMOTE, which can easily be victims of the overlapping noise and overfitting. CTAB-GAN, a specialized tabular GAN, is specifically tailored to deal with the mixed continuous and categorical variables that are common in credit data and has superior generating abilities with imbalanced financial data [20].

In addition to the imbalance of classes, Generative AI is also being applied to macroeconomic stress testing and privacy protection [7]. Privacy laws are very strict on financial data, and they may not be easily shared and utilized across institutions. More recent developments in deep generative modeling, Denoising Diffusion Probabilistic Models (DDPM), are used to generate financial time series [8]. Such diffusion models are able to capture rare, high-impact events and macroeconomic shocks (e.g., market crashes or unexpected changes in interest rates) to generate realistic stress-testing conditions [8]. To achieve high regulatory secrecy, privacy-preserving training (like differentially private stochastic gradient descent, which is also known as DP-SGD) is incorporated into the training of the diffusion model, with the resulting generated synthetic data being temporally faithful and without revealing sensitive user information.

C) Transformer-Based Temporal Sequence Modeling

Evaluating credit risk requires an understanding of sequential, time-series behavioral data. Historically, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks

were the standard for temporal sequence modeling due to their ability to process complex past signals via memory cells and gating mechanisms [1]. However, LSTMs suffer from distinct architectural limitations: their sequential computation prevents parallelization, resulting in slow training times, and they frequently fail to capture very long-term dependencies due to the vanishing gradient problem [1].

To address these constraints, the Transformer architecture has been rapidly adopted in financial modeling. Transformers abandon recurrence entirely, relying instead on multi-head self-attention mechanisms that reduce the distance between any two positions in a sequence to a constant, allowing for parallel data processing and superior extraction of global dependencies [19].

Recent literature demonstrates the superiority of transformers in modeling irregular financial time series for default and fraud prediction. For instance, the Feature Embedded Transformer (FE-Transformer) combines user behavioral sequences with static feature data, achieving significantly higher Area Under the Curve (AUC) and Kolmogorov-Smirnov (KS) metrics than LSTM or XGBoost baselines [19]. In fraud detection, the Graph-Temporal Contrastive Transformer (GTCT) encodes sequential transaction behaviors using a Transformer encoder, addressing irregular time intervals [9]. The embedding output Z_{temp}^{i} for a given account I is formulated as

$$Z_{temp}^{i} = TransformerEncoder(X_i + P_i)$$

where X_i is the sequence matrix and P_i is the positional encoding matrix, ensuring temporal order is preserved. Empirical results show that such transformer-based models effectively capture the evolving dynamics of financial behavior, establishing them as the optimal choice for the dynamic PD/LGD modeling layer of modern risk systems.

D) Cloud-Native Architectures for Financial Risk

The underlying infrastructure should be highly scalable, resilient, and able to orchestrate in real time to be able to operationalize complex AI models in a high-velocity ecosystem. Literature points out the shift from monolithic applications to cloud-native event-driven microservices architectures (EDA) [12]. Institutions attain agility to respond to transactional events in an asynchronous manner by breaking down financial processes into loosely coupled independently deployable services [12]. Interaction among these microservices is based on event brokers (e.g., Apache Kafka or RabbitMQ), which guarantee the reliability of the delivery of messages, event persistence, and sequencing [12].

The most important need of these systems is continuity of operations in case of component failures. Scientists suggest self-healing architectures that use container orchestration systems such as Amazon Elastic Kubernetes Service (EKS) with service meshes such as Istio. This orchestration layer offers active fault detection through telemetry, automatic root cause analysis, and recovery mechanisms. As an example, in case of a crash or spike in latency of a microservice, Kubernetes will automatically restart the failed pods, and Istio will dynamically reroute traffic to stable pods, resulting in zero-downtime deployments. High-availability clusters (e.g., PostgreSQL with Patroni) at the data layer use synchronous replication and automatic failover to guarantee the compliance with ACID and integrity of transactions. Such self-healing cloud-native insurance platforms have been experimentally tested to reduce Mean Time To Recovery (MTTR) by 40 percent and ensure 99.95 percent uptime of their system in the event of a serious failure, making them indispensable to enterprise-level financial risk applications [5].

E) Regulatory Explainability and AI Transparency

While complex machine learning models like Transformers and Gradient Boosting Machines (GBMs) offer superior predictive accuracy, their "black-box" nature fundamentally conflicts with the strict regulatory landscape of the financial sector [18]. Regulators worldwide demand transparency; for instance, the European Union's General Data Protection Regulation (GDPR) dictates that automated decision-making must provide data subjects with meaningful information regarding the logic involved. Similarly, the EU's Artificial Intelligence Act (AIA) establishes stringent requirements for auditability and fairness in high-risk AI systems [5].

Explainable AI (XAI) bridges the gap between high model complexity and regulatory auditability. Post-hoc interpretability frameworks, most notably SHapley Additive exPlanations (SHAP) and Local

Interpretable Model-agnostic Explanations (LIME), are extensively utilized in the literature to unpack GBM and neural network decisions. SHAP, rooted in cooperative game theory, quantifies the marginal contribution of each input feature to a model's prediction [14]. The SHAP value ϕ_J for a feature j is rigorously computed as:

$$\phi_J = \sum_{S \subseteq N/\{j\}} \frac{|S|!\ (|N| - |S|)!}{|N|!} (f(S \cup \{j\}) - f(S))$$

where N is the set of all features, S is a subset of features excluding j, and f(S) is the model's expected output conditioned on subset S [18].

XAI allows financial institutions to audit precise variables by offering both global (those drivers of systemic risk, such as macroeconomic indicators) and local (a justification of the specific loan rejection) explanations [18]. This openness plays a pivotal role in identifying and eliminating discriminatory biases towards protected classes (e.g., age, gender, or ethnicity), procedural fairness, and differential pricing [18]. The literature underlines that XAI integration does not worsen performance, but rather, it is a stabilizing economic mechanism that mitigates risks associated with openness and promotes trust among stakeholders [25].

E) Gap Summary

The literature review indicates that there has been a deep development in the field of technology in individual areas. There is still a major architectural disjunction, however. The reviewed approaches are synthesized in Table 1, shedding light on what has been addressed successfully and what has not been addressed so far in the context of credit risk management on the enterprise-scale.

Table 1
THEMATIC SYNTHESIS OF REVIEWED LITERATURE AND THE ARCHITECTURAL GAP

Technological Domain	**Reviewed Approaches & Capabilities**	**Unresolved Challenges / Research Gaps**
Generative AI	GANs (CTAB-GAN, WGAN) for minority class oversampling. Diffusion models with DP-SGD for privacy-preserving macroeconomic stress testing.	Often studied in isolation on static datasets. Lack of continuous, automated data recalibration pipelines feeding directly into live temporal models.
Temporal Modeling	Transformers (FE-Transformer, GTCT) outperforming LSTMs via parallelized self-attention and handling of long-range behavioral dependencies.	Transformer models are rarely deployed within a self-healing microservices environment that accounts for dynamic, real-time market data streaming.
Cloud-Native Architecture	Event-Driven Microservices (Kafka) and Kubernetes/Istio orchestration enabling 40% MTTR reduction and self-healing fault tolerance.	The high-level orchestration systems frequently operate independently from the AI models, lacking ML-driven resource optimization and automated data drift handling.
Explainable AI (XAI)	SHAP/LIME integration with GBMs, providing regulatory compliance (GDPR/AIA) and local/global feature transparency.	High computational cost of XAI limits real-time dynamic applications. Most studies apply XAI post-hoc to standalone models, not continuous pipelines.

Direct placement of the GDRA As Table 1 demonstrates, existing literature considers Generative AI, transformer-based sequence modeling, cloud-native orchestration, and Explainable AI to be discrete, siloed solutions [8]. No other existing framework has managed to prove the overall cohesion of these separate components to form one system.

The Generative-Dynamic Risk Architecture (GDRA) that is proposed specifically addresses this very gap in integration. The GDRA provides the ability to recalibrate data continuously

and simulate stress scenarios by structurally integrating a cloud-native self-healing data ingestion layer with a GAN/Diffusion-based synthetic oversampling engine. This actively feeds into a Transformer-based temporal modeling layer to make very accurate predictions of PD/LGD, all within an overarching, real-time SHAP interpretability framework. As such, the GDRA will convert these discrete technological accomplishments into a common, enterprise-ready credit risk architecture that naturally fulfills the current requirements to predict with accuracy, infrastructural stability, and strict regulatory disclosure.

III. Methodology

A) Search Strategy and Selection Criteria

In order to develop a strict theoretical and empirical framework of the Generative-Dynamic Risk Architecture (GDRA), the given research paper has carried out a systematic literature review in accordance with the PRISMA 2020 (Preferred Reporting Items to Systematic Reviews and Meta-Analyses) recommendations [15]. The search was conducted in significant academic databases, such as Scopus, IEEE Xplore, ACM Digital Library, and ScienceDirect, containing the literature published between 2015 and 2025 so that it could be relevant to the current time [3].

The search was based on a set of controlled vocabulary words and free-text words in three conceptual categories: (1) predictive modeling (credit risk, default prediction, transformer, temporal sequence); (2) data augmentation (generative AI, GAN, diffusion models, synthetic data); and (3) infrastructure and compliance (microservices, cloud-native, explainable AI, SHAP) [23].

Peer-reviewed scientific journal articles and conference proceedings that offered empirical implementations, architectural frameworks, or performance appraisals of these technologies in financial contexts were strictly included as inclusion criteria [3]. Non-academic literature, discussions of pure theory, not substantiated with empirical research, and works that do not involve particular applications to financial risk or data creation have been filtered out [3]. After automated duplicate elimination, the articles were subjected to title and abstract screening and then to a full-text evaluation to complete the corpus to be synthesized by themes [3].

B) Thematic Synthesis and Comparative Architectural Analysis

The chosen literature was synthesized on the basis of thematic synthesis in order to determine the gaps in the operation of the legacy credit risk systems and to gain the elements that were required in the GDRA. The literature was classified in four critical themes:

1. **Event-Driven Data Infrastructure:** Flexible transaction processing with microservices that are loosely coupled and real-time instead of batch processing [12].
2. **Synthetic Data Generation:** Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPM) are used to address the issue of class imbalance and to generate macroeconomic stress testing [20].
3. **Temporal Sequence Modeling:** A replacement of Recurrent Neural Networks (RNNs/LSTMs) with Transformer-based systems that are able to represent long-range dependencies in irregular financial behavior [19].
4. **Regulatory Interpretability:** The compulsory provision of Explainable AI (XAI) so that sophisticated models can meet the requirements of transparency laws (e.g., GDPR, EU AI Act) [18].

C) Selected Performance Metrics

A set of metrics, which connected predictive classification accuracy, architectural resilience, and interpretability, was chosen to objectively compare the proposed GDRA with legacy baselines.

- **Classification Metrics:** Since the data of credit defaults is inherently imbalanced, threshold-free metrics, including the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the F1-Score, were considered more important than the raw accuracy [9]. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) [19]. These are defined as:

$$TRP = \frac{TP}{TP + FN} \times 100\%$$

$$FPR = \frac{FP}{FP + TN} \times 100\%$$

where TP is True Positives, FN is False Negatives, FP is False Positives, and TN is True Negatives [19]. The Kolmogorov–Smirnov (KS) statistic was also selected to measure the model's capacity to distinguish between default and non-default distributions [19].

- **Architectural Resilience Metrics:** To evaluate the cloud-native orchestration layer, system performance is measured via transaction throughput (Transactions Per Second, TPS) and Mean Time To Recovery (MTTR). MTTR quantifies the self-healing capability of the system during simulated service crashes.

- **Explainability Metrics:** SHapley Additive exPlanations (SHAP) was selected to quantify feature importance [18]. Rooted in cooperative game theory, SHAP provides a unified measure of feature contribution by calculating the marginal impact of a feature across all possible feature subsets [18]. The SHAP value ϕ_j for a feature j is calculated as:

$$\phi_J = \sum_{S \subseteq N/\{j\}} \frac{|S|!\ (|N| - |S|)!}{|N|!} (f(S \cup \{j\}) - f(S))$$

where N is the set of all features, S is a subset excluding j, and f(S) is the model prediction for subset S [18].

D) GDRA Design Rationale

The GDRA was designed as a synthesized solution to the limitations identified in the thematic analysis, structurally integrating data ingestion, augmentation, prediction, and explanation into a unified, continuous pipeline.

Layer 1: Self-Healing Data Architecture

Traditional monolithic batch-processing systems suffer from severe data latency and model drift [12]. Informed by cloud-native computing research, Layer 1 utilizes an event-driven microservices architecture supported by message brokers (e.g., Kafka) and container orchestration platforms like Amazon Elastic Kubernetes Service (EKS) combined with an Istio service mesh. This design choice ensures real-time transaction ingestion, automatic scaling during high-velocity data surges, and "self-healing" capabilities. In the event of a pod crash or latency spike, the orchestration layer automatically restarts failed services and reroutes traffic, ensuring zero-downtime operational continuity.

Layer 2: GAN-based Synthetic Oversampling Engine

To counteract the performance degradation caused by extreme class imbalance in credit default datasets, Layer 2 deploys specialized Generative AI models [20]. Traditional oversampling techniques like SMOTE suffer from overlapping noise; therefore, tabular-specific GANs (such as CTAB-GAN) are utilized to learn the conditional distributions of minority classes and synthesize high-fidelity realistic data. Furthermore, to address macroeconomic volatility, this layer incorporates Denoising Diffusion Probabilistic Models (DDPM) conditioned on stress scenarios (e.g., GDP decline) trained using Differentially Private Stochastic Gradient Descent (DP-SGD) to generate stress-testing datasets that comply with strict financial privacy regulations.

Layer 3: Dynamic PD/LGD Modeling and XAI

The predictive core of the GDRA replaces legacy statistical models and sequential LSTMs with Transformer-based architectures. Transformers, utilizing multi-head self-attention mechanisms, process sequential data in parallel, vastly improving the extraction of long-range behavioral dependencies and irregular time intervals typical of multi-platform borrower activity [9]. To satisfy regulatory requirements for auditability and to prevent algorithmic discrimination, this predictive layer is tightly coupled with a SHAP-based Explainable AI framework. SHAP provides both global interpretability (understanding systemic risk drivers) and local interpretability (generating feature-level justifications for individual loan rejections). This ensures that the highly complex Transformer predictions are fully transparent, equitable, and compliant with regulatory regimes.

IV. Results

A) Thematic Synthesis Findings

The systematic review of the chosen literature supports the concept of the structural inadequacy of the legacy credit risk systems and proves the need for the modernized approach. The reviewed literature can be synthesized thematically,

which demonstrates that four predominant technological pillars are oriented towards the shortcomings of the batch-processed Probability of Default (PD) and Loss Given Default (LGD) models:

1. **Real-Time Processing with Event-Driven Microservices:** The traditional monolithic architecture and batch-processing systems are poorly designed to deal with high-velocity digital financial ecosystems, with extreme latency and model drift [12]. Cloud-native microservices, managed through containerization systems, enable financial institutions to accept streams of transaction events asynchronously, which provides continuity of operations and instant revenue recognition [12].

2. **Generative AI for Imbalanced Data and Stress Testing:** Imbalanced Data and Stress Testing: The datasets of credit card fraud and default are highly imbalanced, with the vast majority of the fraudulent or default cases being rare exceptions [20]. The use of generative adversarial networks (GANs) and diffusion models is both more effective than traditional resampling algorithms (such as SMOTE) at learning the complex, hidden, and non-linear joint distribution of highly imbalanced financial data without adding overlapping noise [20].

3. **Transformers over Recurrent Networks:** Although Long Short-Term Memory (LSTM) networks have been recommended in the past as a preferred model to model sequential behavior, they experience bottlenecks in sequential computation and are unable to model long-term dependencies. The literature reveals that Transformer architectures, based on multi-head self-attention mechanisms, are much more effective at predicting user default risk by picking up irregular time intervals and long-range behavioral dependencies in parallel compared to LSTMs [19].

4. **Regulatory Explainability (XAI):** Complex machine learning algorithms are opaque, which can be a problem when dealing with regulatory systems. The most prominent example is explainable AI (XAI) frameworks, such as SHapley Additive explanations (SHAP), which is a necessary compromise between the predictive accuracy of gradient boosting and transformer models and the transparency mandates of financial regulators [18].

B) GDRA Synthesised Framework

According to the thematic synthesis, we present the Generative-Dynamic Risk Architecture (GDRA), a single, constantly re-calibrating credit risk model. The GDRA integrates the individual successes of cloud computing, generative AI, temporal modeling, and explainable AI into one, enterprise-deployable architecture in the form of three main layers.

Layer 1: Self-Healing Data Architecture

The GDRA is based on a very robust, cloud-native system that is capable of ingesting and processing real-time streams of transactions with as little downtime as possible when a component fails. Based on the cutting-edge cloud deployment strategies, Layer 1 deploys an Amazon Elastic Kubernetes Service (EKS) as a container orchestrator alongside an Istio service mesh [26]. This architecture fully separates the financial processes into autonomous microservices (e.g., billing, auditing, and real-time revenue recognition) interacting asynchronously through event brokers such as Kafka or RabbitMQ [12].

Most importantly, the layer offers a self-healing mechanism that is powered by proactive telemetry. When AWS CloudWatch or Prometheus notices a crash of a service or a spike in latency, Kubernetes will restart the crashed pods, and Istio will automatically reroute network traffic to healthy service endpoints. At the data persistence tier, the GDRA uses a high-availability PostgreSQL cluster operated by Patroni to guarantee synchronous replication and executes automated database failovers. This self-healing design eliminates the latency of the old-fashioned batch processing and makes sure that the downstream modeling layers have continuous and real-time borrower information [12].

Layer 2: GAN-Based Synthetic Oversampling and Stress Engine To resolve the empirical problem of data scarcity, privacy constraints, and highly imbalanced default classes, Layer 2 deploys a sophisticated generative engine. This layer utilizes specialized tabular GANs, such as the Conditional Table GAN (CTAB-GAN), to accurately encode mixed continuous and categorical variables and synthesize high-fidelity minority class data without overfitting the original dataset [20].

Furthermore, to generate robust macroeconomic stress-testing scenarios (e.g., predicting LGD during sudden market volatility), Layer 2 incorporates Denoising Diffusion Probabilistic Models (DDPM). The diffusion model learns to reverse a forward noising process, minimizing the mean-squared error between true noise and the network's prediction:

$$L_{Simple} = E_{X_o,\epsilon\sim N(0,I),t}\,[|\epsilon - \epsilon_0(X_t, t)|^2]$$

where x_0 is the clean financial time series, ϵ is a Gaussian noise vector, and ϵ_θ (x, t) is the neural network's estimate of the noise at diffusion step t [8].

To ensure strict compliance with data protection laws, this layer trains the generative models using Differentially Private Stochastic Gradient Descent (DP-SGD). By applying L2-norm gradient clipping and injecting Gaussian noise during backpropagation, DP-SGD provides mathematical guarantees (e.g., a privacy budget of ϵ =1.2, δ=10^{-5}) that the generated synthetic sequences cannot be reverse-engineered to expose sensitive, individual borrower records [8].

Layer 3: Dynamic PD/LGD Modeling and XAI

The predictive core of the GDRA abandons traditional econometric and static scorecards in favor of Transformer-based temporal sequence modeling. Taking inspiration from the Feature Embedded Transformer (FE-Transformer) and the Graph-Temporal Contrastive Transformer (GTCT), Layer 3 processes the sequential online behavioral data and structural graph relationships of borrower accounts [9].

Instead of processing transaction histories sequentially like an LSTM, the multi-head self-attention mechanism computes temporal relationships in parallel, shrinking the distance between any two transactional events to a constant and capturing long-range dependencies across irregular financial time series [19]. The embedding output for an account I is formalized as: $Z_{temp}^{i} = TransformerEncoder(X_i + P_i)$
where X_i is the transaction sequence matrix and P_i is the positional encoding matrix preserving temporal order.

To satisfy regulatory transparency, the complex, non-linear predictions produced by the Transformer are continuously interpreted using SHapley Additive exPlanations (SHAP). Rooted in cooperative game theory, SHAP provides localized post-hoc explainability by quantifying the exact marginal contribution of each input feature to the final credit decision [18]. The SHAP value ϕ_J is computed as follows:

$$\phi_J = \sum_{S\subseteq N/\{j\}} \frac{|S|!\ (|N| - |S|)!}{|N|!}(f(S\cup\{j\}) - f(S))$$

where N is the total set of features and S is a subset excluding feature j. This ensures that while the GDRA leverages "black-box" predictive power, its outputs remain fully auditable and interpretable for human-in-the-loop oversight.

C) Comparative Performance Analysis

In order to test the synthesised GDRA design, we compared the performance metrics of core elements of the design to traditional legacy baselines checked in the literature. The performance analysis bridges both the architectural resilience (Layer 1) and predictive classification accuracy (Layer 3).

Classification and Predictive Metrics: It has been shown in the literature conclusively that structural modeling, temporal dynamics, and generative data augmentation are dramatically more effective than classical methods. Table 2 summarizes the results of models including Transformer parts (such as the GTCT and FE-Transformer), which are much more effective than Logistic Regression (LR), Random Forests (RF), and single LSTMs in imbalanced credit risk environments [19]. The GTCT also possesses an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.982 and an F1-Score of 0.876, which is an excellent discriminative performance between fraudulent/default and legitimate transactions [9]. Likewise, FE-Transformer produces the highest AUC (0.72) and Kolmogorov-Smirnov (KS) (0.32) statistic in case of integrated behavioral and credit feature data, as compared to AM-LSTM [19].

Table II
COMPARATIVE CLASSIFICATION PERFORMANCE OF PREDICTIVE MODELS

Model Architecture	Accuracy	F1-Score	AUC-ROC
Logistic Regression (Baseline)	0.951	0.358	0.763
Random Forest	0.961	0.576	0.849
LSTM Network	0.962	0.765	0.905
FE-Transformer (GDRA Layer 3)	-	-	0.720 (KS: 0.32)
GTCT (GDRA Layer 3)	0.975	0.876	0.982

Note: Metrics synthesized from distinct empirical evaluations of FE-Transformer and GTCT frameworks.

Ablation studies further prove the necessity of the GDRA's holistic approach: removing the temporal encoder from the GTCT framework caused recall to drop precipitously from 0.867 to 0.743, proving that sequential behavioral cues (like irregular transaction timing) are critical for dynamic risk assessment [9]. Furthermore, when generative models like CTAB-GAN are used for data synthesis prior to training (Layer 2), classifiers demonstrate accuracy improvements of up to 17% on complex datasets by eliminating the overlapping noise that plagues traditional SMOTE resampling [20].

Architectural Resilience Metrics: The shift from batch processing to the GDRA's self-healing cloud-native architecture directly translates into highly optimized operational performance. Empirical evaluations of Kubernetes and PostgreSQL Multi-AZ implementations during simulated infrastructure crashes (e.g., node failures, latency spikes) reveal an operational uptime of 99.95% [26]. The automated fault detection and dynamic rerouting orchestrated by the service mesh reduce the Mean Time To Recovery (MTTR) by 40% compared to manual legacy interventions [26]. Crucially, the load-balancing mechanisms allowed the simulated financial system to increase transaction throughput by 32% even during failure conditions, sustaining volumes of 10,000 transactions per minute [26]. Furthermore, edge-computing integrations in distributed financial architectures demonstrate the capacity to reduce average transaction processing times by 69% [3].

D) Security and Compliance Alignment

The operationalization of any modern credit risk architecture is strictly governed by global financial compliance frameworks (e.g., GDPR, EU Artificial Intelligence Act, PCI-DSS, and SOC2 Type II). The GDRA is fundamentally aligned with these regulatory mandates by design, rather than through post-hoc patching.

First, the integration of DP-SGD in the Generative Engine (Layer 2) ensures formal data privacy. Regulations such as the GDPR demand stringent protections for Personally Identifiable Information (PII) [8]. By incorporating Rényi differential privacy accounting into the DDPM training, the GDRA prevents membership inference attacks and ensures that the synthetic portfolios used to train the Transformers do not leak original borrower histories [8]. The utilization of Zero-Trust security principles, centralized encryption, and robust identity and access management (IAM) at the cloud orchestration layer further satisfies the data-in-transit and data-at-rest requirements mandated by PCI-DSS.

Second, the EU's Artificial Intelligence Act and GDPR dictate that data subjects are entitled to meaningful information regarding the logic of automated decision-making. Black-box transformer models, deployed in isolation, violate these transparency requirements. The GDRA inherently solves this compliance friction through Layer 3's SHAP XAI framework. SHAP provides both global model transparency (e.g., identifying systemic macro-risk drivers) and localized, instance-specific justifications. By quantifying exact feature contributions for a specific rejected credit application, auditors and financial professionals can definitively prove that the AI model did not rely on discriminatory variables (e.g., age, gender, or ethnicity), thereby assuring procedural fairness, eliminating algorithmic bias,

and maintaining rigorous accountability under regulatory scrutiny [5].

V. Discussion

A) Interpretation of Findings

The architectural flaws of legacy credit risk systems as monolithic batch-processing systems with linear modeling assumptions have made them fundamentally incompatible with the real-time pace of digital finance today [12]. The metasynthesis of the literature evidences that these inadequacies can be overcome only when the isolated algorithmic improvements are shifted to a higher level. Instead, the results indicate that there should be a union of three separate capability bundles: Generative AI augments data synthetically, Transformer-based temporal sequence modeling leads to predictive analytics, and cloud-native microservices have a resilient orchestration.

Each of these technologies has individually been very successful in addressing certain bottlenecks in financial risk management. The widespread problems of class imbalance, insufficient data, and strict privacy policies that afflict financial datasets are successfully alleviated by Generative Adversarial Networks (GANs) and diffusion models [8]. GANs can create realistic minority-class images (e.g., fraudulent transactions or loan defaults) by approximating the joint distribution of real-world data, without revealing sensitive personal data [20]. At the same time, Transformer architectures have become the best at managing time by overcoming sequential computation bottlenecks and long-range dependency constraints of conventional Recurrent Neural Networks (RNNs) and LSTMs. Simultaneously, event-driven microservices based on the cloud's deployment have shown the potential to substitute weak monolithic infrastructures, allowing immediate event processing and high resiliency to fault tolerance [12].

The fact that these three clusters have come together to the proposed Generative-Dynamic Risk Architecture (GDRA) is an indication of a radical change of direction in the disciplines of quantitative finance and risk management. This means that credit risk systems in the future should be modeled as self-recalibrating ecosystems, as opposed to a scoring pipeline that is not continually calibrated. With all of these aspects in place, the GDRA makes a practical implementation of the need to move beyond retrospective, batch-processed application scoring to a dynamic, real-time behavioral scoring, driven by the crucial inclusion of Explainable AI (XAI) to assure regulatory compliance [18].

B) Comparison with Existing Frameworks

Compared to the current financial risk frameworks, the GDRA is an important architectural development. Conventional credit scoring is largely based on econometric models, e.g. Logistic Regression, with high interpretability but poorly models the non-linear interactions of multi-platform borrower activity [5]. On the other hand, more recent machine learning applications in finance have often deployed sophisticated algorithms on their own, considering data augmentation, predictive modeling, and system deployment as separate stages.

Indicatively, whereas some models such as the Graph-Temporal Contrastive Transformer (GTCT) have demonstrated impressive predictive accuracy, with an Area Under the Receiver Operating Characteristic Curve (ROC-AUC) of 0.982 and an F1-Score of 0.876 in fraud detection tasks, they are commonly tested in stagnant computational settings as opposed to dynamic enterprise-scale or Likewise, state-of-the-art tabular GANs such as CTAB-GAN offer better generation of imbalanced continuous and categorical variables, but their application in online, automated recalibration pipelines is under-explored [20].

The GDRA distinguishes itself by weaving these components together. Unlike conventional monolithic applications that suffer from substantial data latency and model drift, the GDRA's cloud-native foundation utilizes Kubernetes orchestration, Istio service meshes, and PostgreSQL high-availability clusters to achieve a "self-healing" state [26]. Empirical benchmarks from the literature indicate that such cloud-native configurations can reduce Mean Time To Recovery (MTTR) by 40% and increase transaction throughput by 32% during system failures [26]. Furthermore, the GDRA advances the state of the art by structurally mandating the continuous application of XAI frameworks—specifically SHapley Additive exPlanations (SHAP)—directly onto the Transformer outputs [18]. This circumvents the "black-box" criticism that typically precludes the

deployment of deep neural networks in highly regulated financial environments.

C) Implications

The theoretical and practical implications of the GDRA cut across researchers, industry practitioners, and regulators.

To researchers: These advanced paradigms are used in an integration that reveals key open problems in the field of explainable continual learning. Due to the dynamism in consumer financial activities and macroeconomic factors, the models need to change on a real-time basis. Nonetheless, the existing literature shows that dynamic, real-time XAI delivery of dynamic markets is a computationally demanding problem [18]. Moreover, the use of GANs to do continuous data augmentation presents stability issues, including mode collapse and vanishing gradients, and new studies of adaptive network designs and tailored loss functions are needed [20]. The scholars will also have to look into the dynamic learning of graphs that are dynamic and able to update the relational structures in real time as the new transactions come in and leave behind the assumption of the static graphs that current temporal models do.

To practitioners: The requirements to deploy the GDRA imply a complete transformation of institutional IT infrastructure. Banking institutions need to shift off the old systems to containerized microservice architectures that are event-driven [12]. This migration takes advanced orchestration (e.g., AWS Elastic Kubernetes Service), asynchronous message brokers (e.g., Kafka), and automated fault-detection telemetry [12, 26]. The practitioners also need to be ready to deal with the high computational complexity of training deep generative models and Transformers, which requires high-quality hardware acceleration and efficient resource distribution [18].

To regulators: The paradigm shift to adaptive, AI-based risk architectures presents a challenge to existing regulatory frameworks. Transparency, accountability, and the right to meaningful information about the logic of automated decisions are required by global compliance frameworks like the European Union General Data Protection Regulation (GDPR) and the Artificial Intelligence Act (AIA) [18]. Regulators need to shift to examining the ongoing processes and fairness limits of adaptive AI, rather than the fixed model coefficients. The GDRA offers a blueprint to this, using SHAP and Local Interpretable Model-agnostic Explanations (LIME) to produce local and global feature scores of importance and post-hoc auditing, which can easily and definitively demonstrate the lack of discriminatory bias in decisions based on protected classes [18].

D) Limitations and Threats to Validity

Although the theoretical soundness of the GDRA can be stated, certain limitations and threats to validity should be mentioned. To start with, the GDRA is an architectural design, which is synthesized; it is not an empirically proven system, which is deployed on a large scale in a massive enterprise. The benchmarks of performance described, like the 99% accuracy of LightGBM models with SHAP [18], the 40% decrease in MTTR of self-healing microservices [26], and the high-fidelity of diffusion-driven synthetic data, are based on different, isolated studies in the literature reviewed [25].

Second, there is a definite risk to the ecological validity of the data employed in the studies as a foundation. A large portion of the deep learning and temporal modeling architectures considered work with static, publicly available data (like the UCI Credit Card Default dataset) or synthetic subsets [9]. Such datasets are frequently unable to adequately capture the extreme heterogeneity, non-stationary noise, and fast-changing adversarial behaviour of live, global financial networks [9]. As a result, predictive and generative layers of GDRA can suffer performance loss when subjected to real-world and highly volatile data streams.

Lastly, the variation in regulations is a major practical constraint. The framework presupposes an integrated strategy of AI regulation and algorithmic responsibility, which is strongly shaped by the European and North American regulatory practices [5]. The global financial ecosystem is, however, marked with a strong regulatory fragmentation. This is because universal and standardized thresholds of XAI do not exist in the financial sector: what one jurisdiction may deem as sufficient interpretability to introduce the GDRA may fall short in a different country, making the operationalization of the GDRA across borders difficult [18]. These gaps should be filled in future research by the longitudinal and large-scale industrial validation, as well as by directly

seeking cross-jurisdictional harmonization of regulations.

VI. Conclusion

The digital financial ecosystem structural shift has revealed the underlying weaknesses of the legacy, batch-processed credit risk systems, which are structurally ill-suited to the high-velocity, multi-platform character of the current borrower behavior. This is a key architectural gap that was filled in this study by the systematic examination of the existing technological potentials, followed by the synthesis of the Generative-Dynamic Risk Architecture (GDRA). The proposed GDRA is a structural synthesis of event-driven microservices, generative AI, transformer-based temporal sequence modeling, and Explainable AI (XAI) into a single, coherent architecture by no longer relying on ad hoc algorithmic improvements. As shown by the architectural analysis, cloud-native orchestration platforms, based on containerization and service meshes, can effectively eliminate the data latency intrinsic to traditional monolithic systems and guarantee robust and real-time fault tolerance, as illustrated by a 40% decrease in Mean Time To Recovery (MTTR) during system failures. It is a resilient base that allows streaming transaction streams to be continuously and asynchronously ingested to enable dynamic behavioral scoring.

The empirical synthesis also confirmed the predictive, generative, and compliance superiority, which is based on the core modeling layers of the GDRA. The combination of Generative Adversarial Networks (GANs) and diffusion models was incredibly successful in overcoming the universal problem of extreme class imbalance and strict data privacy, enabling the synthesis of minority-class default events with high fidelity and also simulation of macroeconomic stress scenarios without revealing sensitive consumer data. Based on this strong data augmentation, it was demonstrated that Transformer networks can be greatly outperformed compared to previous recurrent models and that they can use parallelized self-attention mechanisms to learn long-range behavioral dependencies in irregular financial time series. More importantly, the structural integration of post-hoc interpretability methods, namely SHAPley Additive explanations (SHAP), was able to address the so-called black-box dilemma of deep learning models successfully. The analysis proved that the GDRA can be used to mathematically quantify the contribution of exact global and local features, making it clear that highly complex machine learning predictions are fully auditable, free of discriminatory bias, and fully adherent to modern regulatory requirements regarding transparency.

References

[1] Maher Ala'raj et al., "Modelling customers' credit card behaviour using bidirectional LSTM neural networks," Springer, 2021. https://link.springer.com/content/pdf/10.1186/s40537-021-00461-7.pdf

[2] Fadi Thabtah et al., "Data imbalance in classification: Experimental evaluation," ScienceDirect, 2020. https://www.sciencedirect.com/science/article/abs/pii/S0020025519310497

[3] Georgios Lambropoulos et al., "Emerging Technologies in Financial Services: From Virtualization and Cloud Infrastructures to Edge Computing Applications," MDPI, 2026. https://www.mdpi.com/2073-431X/15/1/41

[4] Anil Kumar et al., "Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review," MDPI, 2021. https://www.mdpi.com/2227-9091/9/11/192

[5] Niklas Bussmann et al., "Explainable Machine Learning in Credit Risk Management," Springer, 2021. https://link.springer.com/content/pdf/10.1007/s10614-020-10042-0.pdf

[6] Faisal Ramzan et al., "Generative Adversarial Networks for Synthetic Data Generation in Finance: Evaluating Statistical Similarities and Quality Assessment," MDPI, 2024. https://www.mdpi.com/2673-2688/5/2/35

[7] Nari Park et al., "Synthesizing Individual Consumers′ Credit Historical Data Using Generative Adversarial Networks," MDPI, 2021. https://www.mdpi.com/2076-3417/11/3/1126

[8] Mohammed Hameed Alhameed, "Deep Learning-Based Privacy Preserving Diffusion-Driven Synthetic Finance for Robust Stress Testing," IEEE, 2026. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=11443110

[9] Julius Olaniyan et al., “Graph-Temporal Contrastive Transformer for Financial Fraud Detection Using Transaction Behavior Modeling,” MDPI, 2025. https://www.mdpi.com/1999-4893/18/12/770

[10] Lucas Coelho e Silva et al., “Transformers and attention-based networks in quantitative trading: a comprehensive survey,” https://dl.acm.org/doi/pdf/10.1145/3677052.3698684

[11] Saurabh Kohli, “AI-Driven Orchestration Systems in Cloud-Native Financial Applications: A Framework for Next-Generation Investment Platforms,” Sarcouncil Journal of Engineering and Computer Sciences, 2025. https://sarcouncil.com/download-article/SJECS-477-2025-356-363.pdf

[12] Sravan Komar Reddy Pullamma, “Event-Driven Microservices for Real-Time Revenue Recognition in Cloud-Based Enterprise Applications,” SAMRIDDHI, 2022. https://www.smsjournals.com/index.php/SAMRIDDHI/article/view/3417

[13] Giorgio Visani et al., “Statistical stability indices for LIME: obtaining reliable explanations for machine learning models,” arXiv:2001.11757v2, 2020. https://arxiv.org/pdf/2001.11757

[14] Maryan Rizinski et al., “Ethically Responsible Machine Learning in Fintech,” IEEE Access, 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9869843

[15] Matthew J. Page et al., “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” BMJ, 2021. https://www.bmj.com/content/372/bmj.n71

[16] Toon Calders and Sicco Verwer, “Three naive Bayes approaches for discrimination-free classification," Springer, 2010. https://link.springer.com/content/pdf/10.1007/s10618-010-0190-x.pdf

[17] Mohsen Khosravi et al., “Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews,” Health Services Research and Managerial Epidemiology, 2024. https://journals.sagepub.com/doi/pdf/10.1177/23333928241234863

[18] Satyadhar Joshi, “Gradient Boosting and Explainable AI for Financial Risk Management: A Comprehensive Review,” Preprints, 2025. https://www.preprints.org/frontend/manuscript/9a5116442bda6481224c96533ed8b194/download_pub

[19] Chongren Wang and Zhuoyi Xiao, “A Deep Learning Approach for Credit Scoring Using Feature Embedded Transformer,” MDPI, 2022. https://www.mdpi.com/2076-3417/12/21/10995

[20] Emilija Strelcenia and Simant Prakoonwit, “A Survey on GAN Techniques for Data Augmentation to Address the Imbalanced Data Issues in Credit Card Fraud Detection,” MDPI, 2023. https://www.mdpi.com/2504-4990/5/1/19

[21] Dr. Manish Jain, “Survey Of Resilience Strategies In Cloud Platforms: From Fault Detection To Auto-Recovery,” JGRMA, 2025. https://www.researchgate.net/profile/Manish-Jain-61/publication/396925780_

[22] Hunter Heidenreich et al., “Deconstructing Recurrence, Attention, and Gating: Investigating the transferability of Transformers and Gated Recurrent Neural Networks in forecasting of dynamical systems,” arXiv:2410.02654v1, 2024. https://arxiv.org/pdf/2410.02654

[23] V. Lanzetta, “Transfer learning for financial data predictions: a systematic review,” arXiv:2409.17183, 2024. https://arxiv.org/abs/2409.17183

[24] Arthur Charpentier et al., “Reinforcement Learning in Economics and Finance,” arXiv:2003.10014v1, 2020. https://arxiv.org/pdf/2003.10014

[25] A. S. M. Fahim et al., “Algorithmic Accountability in U.S. Consumer FinTech: Governance Mechanisms for Credit Risk, Fair Lending, and Financial Stability,” Journal of Economics, Finance and Accounting Studies, 2023. https://al-kindipublishers.org/index.php/jefas/article/view/12016

[26] Gowtham Reddy Enjam and Komal Manohar Tekale, “Self-Healing Microservices for Insurance Platforms: A Fault-Tolerant Architecture Using AWS and PostgreSQL,” International Journal of AI, Big Data, Computational, and Management Studies, 2024. https://ijaibdcms.org/index.php/ijaibdcms/article/view/252