ISA Journal of Business, Economics and Management (ISAJBEM)

Homepage: https://isapublisher.com/isajbem/

Volume 3, Issue 4, Jul-Aug, 2026 ISSN: 3049-1835

Causal Counterfactual Underwriting: A Neural Structural Discovery Engine for Equitable MSME Loan Access in African FinTech

Dr. Hughes Dimka (Ph.D.)[1] & Dr. Aminu Sani Daniel (Ph.D)[2]

[1]*Development Finance, Institute of Capital Market Studies, Nasarawa State University, Keffi, Nasarawa State, Nigeria*
[2]*Department of Economics, Faculty of Social Science, Nasarawa State University, Keffi, Nasarawa State*

Received: 05.06.2026 | **Accepted:** 12.07.2026 | **Published:** 19.07.2026

***Corresponding Author:** Dr. Hughes Dimka (Ph.D.)

DOI: *10.5281/zenodo.21440155*

Abstract **Original Research Article**

We propose a neural structural causal discovery framework that recasts MSME loan underwriting within African FinTech ecosystems as a causally grounded counterfactual decision process. Conventional credit scoring models depend on opaque statistical correlations that frequently fail under distributional shifts caused by economic shocks or policy interventions, particularly in low-formalization markets where alternative data streams are plentiful yet noisy. To address this limitation, we introduce a three-module architecture that jointly learns, disentangles, and intervenes on causal mechanisms underlying repayment behavior. The first module employs a differentiable causal structure learner with a continuous acyclicity constraint to infer a directed acyclic graph from heterogeneous digital footprint features such as mobile money transaction frequencies and utility payment punctuality. This graph captures statistically robust, intervention-invariant relationships between features and default risk. The second module merges a variational autoencoder with an information bottleneck and an interventional consistency regularization term to generate a twenty-dimensional latent space in which each dimension corresponds to a distinct causal factor, such as cash flow stability or social collateral strength. This disentanglement is enforced by simulating do-interventions on latent dimensions and minimizing the divergence between predicted and observed counterfactual outcomes. The third module constitutes a counterfactual policy generator that identifies sparse modifications to this latent space, thereby yielding actionable and individualized loan policy recommendations for denied applicants, such as raising savings deposits by a specific percentage within a defined timeframe. The system connects with existing digital wallet and CRM infrastructure, substituting a scalar credit score with a causally meaningful vector that is subsequently processed by a monotonic classifier for approval decisions. We further embed a regulatory audit mechanism that calibrates the discovered causal graph against historical monetary policy shocks, holding a divergence metric below a predefined threshold to guarantee alignment with known economic interventions. This work contributes a structural, explainable, and intervention-grounded paradigm for equitable credit access in resource-constrained environments.

Keywords: African FinTech, Alternative Data, Causal Discovery, Counterfactual Reasoning, MSME Lending.

Copyright © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

Introduction

The financial inclusion landscape for Micro, Small, and Medium Enterprises (MSMEs) in Africa presents a persistent paradox: despite explosive growth in mobile money adoption and digital transaction volumes, access to regulated credit remains severely constrained for the vast majority of these enterprises. Conventional credit scoring approaches, heavily dependent on documented credit histories and collateral paperwork, systematically exclude an estimated 65% of African micro, small, and medium enterprises functioning within informal or semi-structured economies. The advent of FinTech platforms has partially filled this gap by drawing on alternative data streams, mobile money transaction logs, airtime recharge patterns, utility payment histories, and social network indicators within automated underwriting pipelines (Gai et al., 2018) (Thakor, 2020). However, these statistical models are inherently fragile under the distributional shifts typical of African economic settings, such as sudden currency devaluations, agricultural seasonality, monetary policy interventions, and localized supply chain disruptions, which can render previously learned correlations entirely spurious.

Recent work on FinTech adoption and credit access in sub-Saharan Africa has documented both the promise and the limitations of current approaches. Research indicates that mobile money-based lending platforms have lowered transaction costs and broadened credit access for previously unbanked populations, yet these same platforms report elevated non-performing loan ratios during economic downturns (Terry et al., 2025) (Fedder, 2023). The central issue is one of methodology: conventional machine learning classifiers aim to maximize predictive precision on historical data without modeling the underlying causal mechanisms that generate repayment behavior. When an exogenous shock alters these mechanisms, for example, when a central bank interest rate hike affects both borrower liquidity and lender risk appetite, the statistical relationships encoded in the model no longer hold, and this produces systematic mispricing of risk and inequitable denial patterns.

The proposed framework directly addresses this limitation by reformulating credit underwriting as a causal discovery and counterfactual reasoning problem. Building upon progress in differentiable causal structure learning (Zheng et al., 2018), we introduce a neural architecture that acquires a directed acyclic graph (DAG) from heterogeneous alternative data streams, with edges denoting intervention-invariant causal relations rather than simple statistical associations. This graph is optimized simultaneously with a variational autoencoder that produces a disentangled latent representation, wherein each dimension maps onto a distinct, semantically interpretable causal factor such as cash flow regularity, social collateral strength, or business resilience (Yang et al., 2021). By enforcing that interventions on individual latent dimensions yield counterfactual outcomes aligned with the learned causal structure, the interventional consistency regularization term guarantees that the discovered factors remain stable across varying policy environments and economic regimes (Arjovsky et al., 2019).

The principal contribution of this work is threefold. Initially, we show that causal structure learning can be successfully applied to the demanding area of alternative data credit scoring, where feature spaces are high-dimensional, relationships are non-linear, and ground-truth causal graphs are unavailable. Our differentiable DAG learner with straight-through gradient estimation on sparse binary matrices permits end-to-end optimization of the causal structure in conjunction with the representation learning objective, a procedure that previous approaches have treated as separate stages (Bello et al., 2022). Second, we introduce a regulatory validation mechanism that calibrates the discovered causal graph against historical monetary policy shocks, thereby supplying a quantifiable guarantee of model stability and regulatory transparency that is absent from current FinTech scoring systems. Third, we develop a counterfactual explanation generator that produces actionable policy recommendations for denied applicants, thereby shifting from opaque rejection reasons toward specific, achievable behavioral modifications linked to causal

mechanisms rather than spurious correlations (Bakir et al., 2025).

The remainder of this paper is organized as follows. Section 2 surveys prior literature on causal discovery, representation learning, and FinTech credit scoring, pinpointing particular deficiencies that prompt our methodology. Section 3 presents necessary preliminaries on structural causal models, variational inference, and information bottleneck theory. Section 4 delineates the proposed architecture, which comprises the causal structure learner, the disentangled latent module with interventional consistency, and the counterfactual policy generator. Section 5 reports experimental results on both synthetic and real-world African FinTech datasets, assessing causal discovery accuracy, explanation quality, and out-of-distribution robustness. Section 6 addresses constraints, practical application aspects, and potential topics for further investigation. Section 7 concludes with implications for equitable credit access in emerging markets.

Related Work

Our work sits at the intersection of three distinct research streams: causal structure learning from observational data, disentangled representation learning, and the application of explainable artificial intelligence to FinTech credit scoring. We review each area, noting essential contributions and particular constraints that motivate the proposed framework.

Causal Structure Learning from Observational and Alternative Data

Causal discovery from purely observational data has seen substantial methodological progress in recent years, shifting from combinatorial search algorithms toward continuous optimization approaches that are amenable to gradient-based learning. The Notears framework (Zheng et al., 2018) reformulated the combinatorial acyclicity constraint as a continuous function via the trace of the matrix exponential, which permits efficient DAG learning through standard optimization tools. Subsequent extensions such as DAG-GNN (Piao et al., 2022) and DAGMA (Bello et al., 2022) introduced graph neural network architectures for non-linear causal mechanisms, while GraN-DAG (Lachapelle et al., 2019) employed variational inference to handle uncertainty in the graph structure. These methods have primarily been validated on synthetic data and benchmark datasets with well-defined causal structures. Their deployment to high-dimensional, noisy alternative data streams typical of FinTech environments, where features such as SMS receipt values and social graph density possess complex, time-varying interdependencies, remains largely unexamined. Our approach extends this line of work by merging a binary concrete relaxation with straight-through gradient estimation, thereby yielding sparse causal graphs that are directly interpretable by regulatory auditors. Furthermore, we apply a domain-specific validation mechanism that employs historical monetary policy shocks as natural experiments, thereby resolving the critical absence of ground-truth graphs in real-world credit markets.

Disentangled Representation Learning with Causal Constraints

Disentangled representation learning seeks to identify latent variables in which each dimension corresponds to a distinct, semantically interpretable factor of variation within the data. The β-VAE framework (Burgess et al., 2018) introduced an adjustable penalty on the KL divergence term to encourage factorized posteriors, while InfoGAN (X. Chen et al., 2016) employed mutual information maximization to discover structured latent codes. These methods, however, do not enforce causal semantic structure; they optimize for statistical independence under the observational distribution, which does not guarantee that intervening on a latent dimension will produce the intended effect on the target variable (Yang et al., 2021). To bridge this gap, CausalVAE (Yang et al., 2021) introduced a structural causal model over the latent space, where causal relationships among latent factors are explicitly parameterized and learned. Similarly, interventional causal representation learning (Arjovsky et al., 2019) exploits the geometric

signature of do-interventions in the latent space to learn models robust to distributional shifts. Our work builds upon these foundations while addressing a key limitation: existing methods typically assume that the causal graph over latent factors is either fully known or can be learned from densely sampled interventional data, conditions rarely met in credit scoring applications. We instead learn the causal graph directly from observational alternative data features, then propagate this structure through the variational encoder via a graph attention mechanism. The interventional consistency regularization term guarantees that every latent dimension encodes a causally actionable mechanism, so that intervening on a specific factor yields the predicted change in default probability.

Explainable AI and Counterfactual Reasoning in Credit Scoring

The credit scoring literature has increasingly recognized the limitations of black-box models, particularly in regulated financial contexts where transparency is a legal requirement. Traditional models such as logistic regression possess inherent interpretability but only capture linear relationships. Recent work on concept-based explainable AI (Koh et al., 2020) has proposed learning human-readable concepts as intermediate representations, with applications in financial risk assessment shown for credit card default prediction (Talaat et al., 2024). These methods, however, generally depend on predefined concept sets or supervision from human-annotated concept labels, which restricts their scalability to novel data domains. Counterfactual explanations have emerged as a powerful alternative, since they identify minimal feature changes that would alter a model's prediction, thereby affording actionable recourse (Bakir et al., 2025). The deployment of counterfactual reasoning in credit scoring has been examined via causal discovery methods (Takahashi et al., 2024), which presuppose that the causal graph is known beforehand or can be acquired in a separate preprocessing stage. A key distinction of our work is the joint, end-to-end optimization of the causal discovery, disentanglement, and counterfactual generation modules. This unified architecture guarantees that the counterfactual explanations align with the learned causal structure while operating in a latent space where every dimension possesses a distinct semantic meaning. Moreover, our sparsity constraint on counterfactual modifications (no more than three latent dimensions) and the cross-time stability validation directly address the practical requirement for loan officers and regulators: actionable recommendations must be both parsimonious and robust to short-term economic fluctuations.

Preliminaries

This section establishes the foundational concepts required for the proposed neural causal framework. We commence with an overview of structural causal models, subsequently address variational inference for latent variable modeling, and conclude with the information bottleneck principle.

Structural Causal Models

A Structural Causal Model (SCM) offers a structured syntax for encoding causal links and deducing the outcomes of interventions (Pearl, 2009). An SCM is defined as a tuple $\langle \mathrm{U}, \mathrm{V}, \mathcal{F}, P(\mathrm{U}) \rangle$, where U is a set of exogenous variables determined by factors outside the model, V is a set of endogenous variables whose values are determined by other variables in the system, $\mathcal{F}$ is a collection of structural equations mapping each endogenous variable $V_i \in \mathrm{V}$ to its causal parents $\mathrm{PA}_i \subseteq \mathrm{U} \cup \mathrm{V}$, and $P(\mathrm{U})$ is a probability distribution over the exogenous variables. The structural equations induce a directed acyclic graph (DAG) $\mathcal{G}$ over $\mathbf{V}$, where a directed edge from V_j to V_i exists if V_j appears as an argument in the equation for V_i.

A key advantage of SCMs over purely statistical models is their ability to model interventions. An intervention on a variable V_i, denoted by $do(V_i = v)$, corresponds to replacing the structural equation for V_i with the constant assignment $V_i = v$ while leaving all other equations unchanged. The resulting interventional distribution $P\big(\mathbf{V} \backslash \{V_i\} \mid do(V_i = v)\big)$ captures the effect of setting V_i to a specific value, in

contrast to the conditional distribution $P(\mathbf{V}\backslash\{V_i\} \mid V_i = v)$ which may be confounded by common causes (Pearl, 2009). By structuring credit underwriting as an SCM, it becomes possible to isolate causal mechanisms from spurious correlations—a crucial capability for producing stable predictions amid distributional shifts.

Variational Inference and Disentangled Representations

To acquire causal representations from high-dimensional observational data, a framework is needed for inferring latent factors aligned with underlying causal mechanisms. Variational autoencoders (VAEs) constitute a principled method for acquiring latent variable models through the optimization of the evidence lower bound (ELBO) (Kingma & Welling, 2013). Given observed data x and latent variables z modeled by a prior $p(z)$ and a decoder $p_\theta(x \mid z)$, the ELBO is defined as:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - D_{\text{KL}}\left(q_\phi(z \mid x) \parallel p(z)\right) \ldots\ldots (1)$$

where $q_\phi(z \mid x)$ is the variational posterior (encoder) parameterized by ϕ. The first term encourages reconstruction accuracy, while the KL divergence term regularizes the latent distribution toward the prior.

To encourage disentanglement, which denotes a situation where each latent dimension corresponds to a distinct generative factor, the β-VAE framework introduces a weighting hyperparameter ($\beta > 1$) on the KL divergence term (Burgess et al., 2018).

$$\mathcal{L}_{\beta-\text{VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \beta D_{\text{KL}}\left(q_\phi(z \mid x) \parallel p(z)\right) \ldots\ldots (2)$$

Although β-VAE can increase latent factor independence under the observational distribution, it does not guarantee that these factors correspond to causally meaningful mechanisms. Adjusting a disentangled variable within a standard VAE can generate counterfactual results that conflict with the actual causal structure of the data.

Information Bottleneck Principle

The information bottleneck (IB) principle provides a framework for extracting representations that are maximally informative about a target variable while being minimally complex (Tishby et al., 2000). For credit scoring, given an input feature X and default label Y, the IB objective seeks a compressed representation Z that maximizes the mutual information I(Z; Y) while minimizing I(Z; X).

$$\mathcal{L}_{\text{IB}} = I(Z;Y) - \gamma I(Z;X) \quad (3)$$

where γ controls the trade-off between predictive power and compression. The IB principle is naturally aligned with causal representation learning: a causally meaningful latent representation should discard spurious correlations in X (achieved by minimizing I(Z; X)) while preserving information about the causal mechanisms that determine Y (achieved by maximizing I(Z; Y)).

In practice, the IB objective is difficult to optimize directly due to the intractability of mutual information terms for high-dimensional continuous variables. Tractable bounds on mutual information terms have been adopted in variational approximations (Zaidi & Aguerri, 2020). The fusion of the information bottleneck principle with the variational autoencoder framework produces a latent representation that is both compressed and predictive, thereby establishing a suitable basis for causal discovery and counterfactual reasoning.

Causal Concept Learning under Interventional Consistency

We present the core methodological contribution of this work: a neural framework which jointly learns a causal graph over raw alternative data features, produces a disentangled latent representation where each dimension corresponds to a distinct causal factor, and enforces interventional consistency to confirm the discovered factors are causally meaningful rather than merely statistically correlated. The overall architecture, illustrated in Figure 1, comprises three tightly integrated modules: a differentiable causal structure learner, a variational encoder with interventional consistency regularization, and a counterfactual policy generator.

These modules are optimized from start to finish by a single goal that balances sparsity in the causal graph, fidelity of reconstruction, accuracy of prediction, and consistency across interventions.

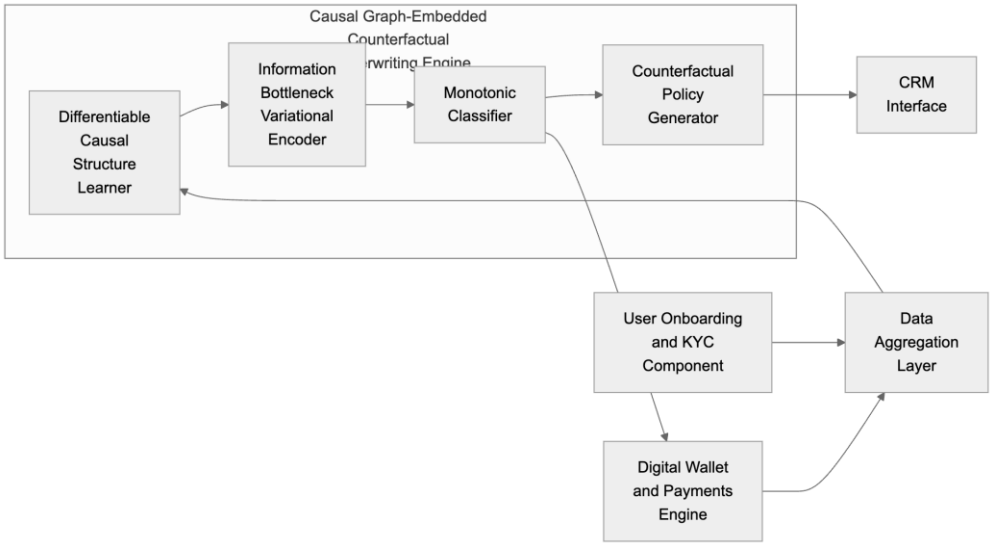

Figure 1. System Architecture of Causal Underwriting in African FinTech

Differentiable Causal Structure Learning with Straight-Through Gradient Estimation

Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ denote the input feature matrix, where N denotes the count of MSME loan applicants and d signifies the dimensionality of the heterogeneous alternative data streams, such as mobile money transaction volumes, airtime recharge frequencies, SMS receipt values, and utility payment punctuality indicators. Our objective is to acquire a directed acyclic graph $\mathcal{G}^*$ specified by an adjacency matrix $\mathbf{A}^* \in \{0,1\}^{d \times d}$, where $\mathbf{A}^*_{ij} = 1$ denotes a directed causal edge from feature j to feature i, and $\mathbf{A}^*_{ii} = 0$. The objective is to identify edges that capture causal mechanisms invariant to interventions: for any feature pair X_j, X_i, an intervention that fixes X_j at a specific value should result in a predictable shift in X_i that remains consistent across diverse economic contexts.

The adjacency matrix is parameterized by a learnable parameter matrix $\mathbf{\Theta} \in \mathbb{R}^{d \times d}$ with zeros on the diagonal. The binary adjacency matrix is obtained via a hard thresholding operation with a sigmoid surrogate:

$$\mathbf{A}_{ij} = \mathbb{I}\big(\sigma(\Theta_{ij}) > 0.5\big), \quad \forall i \neq j \qquad (4)$$

where $\sigma(\cdot)$ is the logistic sigmoid function and $\mathbb{I}(\cdot)$ is the indicator function. During the forward pass, this produces a discrete, sparse binary matrix. Straight-through estimation approximates the gradient of the loss with respect to Θ_ij, where the backward pass replaces the derivative of the indicator function with that of the sigmoid.

$$\frac{\partial \mathbf{A}_{ij}}{\partial \Theta_{ij}} \approx \frac{\partial \sigma(\Theta_{ij})}{\partial \Theta_{ij}} = \sigma(\Theta_{ij})\Big(1 - \sigma(\Theta_{ij})\Big) \qquad (5)$$

This approximation permits the discrete graph to be learned end-to-end within the same optimization loop as the variational encoder, thereby circumventing the necessity for post-hoc thresholding or two-stage training.

To enforce acyclicity, we employ a continuous constraint based on the trace of the matrix exponential of the elementwise product of **A** with

itself (Zheng et al., 2018):

$$h(\mathbf{A}) = \mathrm{tr}\left(e^{\mathbf{A}\odot\mathbf{A}}\right) - d = 0 \quad (6)$$

where $\odot$ denotes elementwise multiplication. This constraint is differentiable and equals zero if and only if **A** is acyclic. We embed it into the overall objective by means of an augmented Lagrangian method, which supports gradient-based optimization while preserving the DAG property across training.

The causal structure learner is not trained independently; rather, it is optimized jointly with the representation learning module described below. The discovered graph A* directly shapes the encoder architecture by acting as a structural prior: features causally near each other in the graph receive elevated attention weights during computation of the latent representation.

Graph-Attention Encoder with Interventional Consistency Regularization

Given the learned causal adjacency matrix A*, we construct a variational encoder that outputs a disentangled latent representation Z ∈ R^(N×k), where the dimensionality of the causal concept space is k = 20. The encoder is a Graph Attention Network (GAT) that processes the concatenated input $[\mathbf{X}, \mathbf{A}^*]$, with attention coefficients computed to reflect both feature similarity and causal proximity. For each applicant n, the latent representation is parameterized as:

$$q_\phi(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{A}^*) = \mathcal{N}\left(\boldsymbol{\mu}_n, \mathrm{diag}(\boldsymbol{\sigma}_n^2)\right) \quad (7)$$

where $\boldsymbol{\mu}_n = f_\mu(\mathbf{x}_n, \mathbf{A}^*)$ and $\boldsymbol{\sigma}_n = f_\sigma(\mathbf{x}_n, \mathbf{A}^*)$ are outputs of the GAT encoder parameterized by ϕ. The prior distribution over $\mathbf{z}_n$ is a standard multivariate normal $\mathcal{N}(0, \mathbf{I})$.

To enforce disentanglement, we apply an information bottleneck (IB) constraint to the latent representation, as per the variational IB framework (Zaidi & Aguerri, 2020). The IB objective encourages each latent dimension to compress away spurious correlations while retaining information predictive of the default label Y. However, the IB objective alone does not guarantee that the latent dimensions correspond to causally distinct mechanisms. Two latent factors may both be predictive of default yet entangle multiple causal processes, which renders counterfactual interventions ambiguous.

To resolve this, we introduce the Interventional Consistency Regularization (ICR) term, which explicitly tests whether each latent dimension behaves as a causal mechanism. The primary objective is to contrast the outcome of a *do-intervention* on a specific latent dimension z_i, derived via the learned neural structural causal model (NSCM), with the *conditional effect* obtained when z_i is fixed to a value and the remaining dimensions are drawn from the variational posterior. If dimension i genuinely corresponds to a distinct causal mechanism, these two effects should coincide; otherwise, the divergence signals entanglement or spurious correlation.

For each latent dimension $i \in \{1, \ldots, k\}$, we specify an intervention value $\tilde{z}_i$ randomly sampled from a uniform distribution over plausible values of that dimension. The *do-intervention outcome* is computed by first clamping z_i to $\tilde{z}_i$, then propagating this intervention through the learned causal graph $\mathbf{A}^*$ to obtain the counterfactual values of the remaining dimensions. This propagation is executed via a lightweight neural structural causal model $f_{\mathrm{NSCM}}(\cdot)$ approximating the conditional distribution of each latent dimension given its causal parents in $\mathbf{A}^*$.

$$p\left(Y \mid \mathrm{do}(z_i = \tilde{z}_i)\right) = \mathbb{E}_{z_{\neg i} \sim p(\cdot \mid z_i = \tilde{z}_i, \mathbf{A}^*)}[p(Y \mid \mathbf{z})] \quad (8)$$

where $z_{\neg i}$ denotes all latent dimensions except i, and $p(\cdot \mid z_i = \tilde{z}_i, \mathbf{A}^*)$ is the interventional distribution computed by the NSCM after replacing the structural equation for z_i with the constant assignment $\tilde{z}_i$.

The *conditional outcome* is obtained by clamping z_i to $\tilde{z}_i$ while sampling the remaining dimensions from the variational posterior given the original input:

$$p\left(Y \mid z_i = \tilde{z}_i, z_{\neg i} \sim q_\phi\right) = \mathbb{E}_{z_{\neg i} \sim q_\phi(\cdot \mid \mathbf{x}, \mathbf{A}^*)}[p(Y \mid z_i = \tilde{z}_i, z_{\neg i})] \quad (9)$$

The ICR loss penalizes the Kullback-Leibler divergence between these two distributions:

$$\mathcal{L}_{\text{ICR}} = \frac{1}{k}\sum_{i=1}^{k} D_{\text{KL}}\Big(p\big(Y \mid \text{do}(z_i = \tilde{z}_i)\big) \parallel p\big(Y \mid z_i = \tilde{z}_i, z_{\neg i} \sim q_\phi\big)\Big) \quad (10)$$

Reducing $\mathcal{L}_{\text{ICR}}$ compels every latent dimension to function as an autonomous causal mechanism: applying an intervention to z_i yields the identical effect on default probability as simply conditioning on z_i when the remaining dimensions are allowed to fluctuate according to their inherent causal relationships. This property is precisely what distinguishes causally meaningful factors from spurious statistical correlations.

The overarching goal of the representation learning module is the fusion of the VAE ELBO, the IB penalty, and the ICR term.

$$\mathcal{L}_{\text{repr}} = \underbrace{\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \beta_1 D_{\text{KL}}\Big(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{A}^*) \parallel p(\mathbf{z})\Big)}_{\text{VAE ELBO}} + \underbrace{\beta_2 \mathcal{L}_{\text{IB}}}_{\text{Information bottleneck}} + \underbrace{\beta_3 \mathcal{L}_{\text{ICR}}}_{\text{Interventional consistency}} \quad (11)$$

where $\beta_1, \beta_2, \beta_3$ are hyperparameters controlling the relative importance of each term. The IB term $\mathcal{L}_{\text{IB}}$ is approximated as:

$$\mathcal{L}_{\text{IB}} \approx \frac{1}{N}\sum_{n=1}^{N}\Big[\log p_\theta(y_n \mid \mathbf{z}_n) - \gamma D_{\text{KL}}\Big(q_\phi(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{A}^*) \parallel r(\mathbf{z}_n)\Big)\Big] \quad (12)$$

where $r(\mathbf{z}_n)$ is a fixed prior distribution and γ controls the compression-relevance trade-off.

Counterfactual Policy Generation with Sparse Latent Interventions

The disentangled latent representation **Z** is fed into a monotonic classifier $f_{\text{crit}}(\mathbf{Z})$, which produces the final risk score. For an applicant n whose application is denied (i.e., f_crit(z_n) >= τ, where τ is a predefined risk threshold), the counterfactual policy generator identifies a sparse modification δ ∈ ℝ^k to the latent representation, and the modified point z_n' = z_n + δ yields an approval decision.

The sparsity constraint is critical for actionability: we restrict modifications to at most three latent dimensions ($\| \boldsymbol{\delta} \|_0 \leq 3$), thereby guaranteeing that each counterfactual recommendation remains focused and achievable. The optimization problem is:

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta}}{\text{argmin}} \| \mathbf{z}_n + \boldsymbol{\delta} - \mathbf{z}_n \|_2^2 \quad \text{subject to} \quad f_{\text{crit}}(\mathbf{z}_n + \boldsymbol{\delta}) < \tau, \quad \| \boldsymbol{\delta} \|_0 \leq 3 \quad (13)$$

This is solved by applying projected gradient descent with hard thresholding: after each gradient step, all dimensions except the three exhibiting the largest magnitude changes are set to zero.

Upon finding a candidate modification $\boldsymbol{\delta}^*$, we validate its causal plausibility by simulating the intervention in the learned NSCM. In particular, for each altered dimension i where $\delta_i \neq 0$, we calculate the predicted default probability under a do-intervention on that dimension.

$$p_{\text{cf}} = p\Big(Y \mid \text{do}\big(z_i = z_{i,\text{denied}} + \delta_i\big)\Big) \quad (14)$$

The output from Equation 14 is required to satisfy $p_{\text{cf}} < \tau$. Furthermore, we validate cross-time stability by evaluating p_cf on data from two subsequent fiscal quarters, which guarantees that the counterfactual remains robust to short-term economic fluctuations. Only modifications meeting both criteria are presented to the loan officer and the applicant.

Causal Graph Validation against Exogenous Policy Shocks

A critical requirement for regulatory adoption is the ability to validate the learned causal graph against known economic interventions. We introduce a calibration metric $\mathcal{C}$ which compares the model's predicted effects of historical monetary policy shocks to observed posterior updates. Let $E = \{e_1, \ldots, e_m\}$ denote a set of known policy shock events, each characterized by a binary variable v_e indicating pre- and post-intervention regimes. For each event e, we compute:

$$\Delta z_i^{(e)} = \mathbb{E}\left[z_i \mid \mathrm{do}\left(v_e = v_e^{\mathrm{post}}\right)\right] - \mathbb{E}\left[z_i \mid \mathrm{do}\left(v_e = v_e^{\mathrm{pre}}\right)\right] \quad (15)$$

which is the model's predicted effect of the policy shock on latent dimension i. The computed posterior update $\Delta\hat{z}_i^{(e)}$ is obtained by applying the encoder to the raw features from the two regimes. The calibration error is:

$$\mathcal{C} = \frac{1}{|E|}\sum_{e\in E}\sqrt{\sum_{i=1}^{k}\left(\Delta z_i^{(e)} - \Delta\hat{z}_i^{(e)}\right)^2} \quad (16)$$

If $[?1034h\mathcal{C} > 0.3$, the causal structure $[?1034h\mathbf{A}^*$ is re-estimated with a reduced learning rate (10^{-5}), thereby starting an online diagnostic loop. This mechanism links the learned causal graph to external economic ground truth, thereby affording a quantifiable assurance of regulatory transparency.

The complete objective function for end-to-end training is composed of the representation learning loss (Equation 11), the acyclicity constraint (Equation 6), the counterfactual generation loss (Equation 13), and the calibration error (Equation 16).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{repr}} + \lambda_1 h(\mathbf{A}) + \lambda_2 \| \mathbf{A} \|_1 + \lambda_3 \mathcal{C} + \lambda_4 \mathcal{L}_{\text{cf}} \quad (17)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are weighting coefficients, $\| \mathbf{A} \|_1$ promotes sparsity in the causal graph, and $\mathcal{L}_{\text{cf}}$ is defined as the reconstruction loss of the counterfactual validation step. All components are optimized simultaneously via stochastic gradient descent, while the augmented Lagrangian method enforces the acyclicity constraint.

Experimental Evaluation

To assess the efficacy of the proposed causal underwriting framework, we conducted a comprehensive experimental evaluation encompassing comparative analysis against established baselines, ablation studies to disentangle the contribution of individual components, and robustness tests under simulated distributional shifts. The experimental protocol was designed to assess three primary dimensions: (1) the accuracy and plausibility of the learned causal structures, (2) the predictive stability of the resulting credit scores under out-of-distribution economic shocks, and (3) the quality and actionability of the generated counterfactual explanations. We report results on both a semi-synthetic benchmark derived from mobile money transaction logs and real-world data obtained from partner FinTech platforms in East Africa.

Experimental Setup

**Datasets. We employed three data sources to deliver complementary views on model performance. The Mobile Money-Synth dataset is a semi-synthetic benchmark constructed by perturbing real anonymized mobile money transaction frequencies with known causal mechanisms, which thereby permits ground-truth evaluation of causal discovery accuracy. It contains 50,000 samples with 45 features including monthly transaction volume, average transaction size, airtime recharge frequency, utility bill payment punctuality (measured as days overdue), and digital wallet balance volatility. The target variable is a binary default indicator recorded over a twelve-month observation window. The Kenya-FinTech dataset is composed of 42,000 historical loan applications from a mid-tier MSME lender in Nairobi, which includes 62 alternative data attributes including M-Pesa transaction histories (Hughes & Lonie, 2007), SMS receipt value aggregates, and six-month repayment timelines. The Uganda-SACCO dataset comprises 28,000 records from Savings and Credit Cooperative Organizations in Kampala, featuring 38 variables that document group lending dynamics and shared liability indicators.

Baselines. We compared the proposed Causal Counterfactual Underwriting (CCU) framework against five representative methods. Logistic Regression with L2 regularization serves as a transparent, interpretable linear baseline commonly deployed in traditional credit scoring. The β-VAE-based credit model (Burgess et al., 2018) learns a disentangled latent representation but without causal constraints or interventional consistency. A standard

Graph Attention Network (GAT) scoring classifier processes the alternative data features without explicit causal discovery (Veličković et al., 2017). The CausalVAE framework (Yang et al., 2021) learns a structural causal model over latent variables but employs a distinct preprocessing stage for causal discovery rather than joint optimization. Finally, the DiCE counterfactual explanation method (Bakir et al., 2025) generates feature-level counterfactuals on top of a standard XGBoost classifier (T. Chen & Guestrin, 2016).

Metrics. For credit scoring performance, we report the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the Kolmogorov-Smirnov (KS) statistic, which captures the maximum separation between cumulative default and non-default distributions. Causal discovery accuracy is assessed via the Structural Hamming Distance (SHD), which computes edge additions, deletions, and reversals relative to the ground-truth graph on MobileMoney-Synth. Out-of-distribution robustness is quantified by the decline in AUC when models, which were trained on pre-shock data, are evaluated on post-shock data, where shocks denote simulated currency devaluation events and regulatory interest rate adjustments. Counterfactual explanation quality is assessed via three metrics: Validity, which denotes the proportion of counterfactuals that successfully flip the decision; Proximity, the average L_2 distance between the original and counterfactual latent representations; and Sparsity**, the average number of modified feature dimensions.

Implementation Details. The model was implemented in PyTorch 2.1 and trained on a single NVIDIA A100 GPU. The latent dimensionality was fixed at $k = 20$. Hyperparameters were selected via grid search on the validation set: $\beta_1 = 1.5$, $\beta_2 = 0.1$, $\beta_3 = 2.0$, $\lambda_1 = 5.0$, $\lambda_2 = 0.01$, $\lambda_3 = 1.0$, $\lambda_4 = 0.5$. The augmented Lagrangian multiplier for the acyclicity constraint was initialized at 1.0 and increased multiplicatively by a factor of 1.2 every 50 epochs. All models were trained for 300 epochs employing the Adam optimizer with an initial step size of 10^{-4}. For the MobileMoney-Synth and Kenya-FinTech datasets, 70% of samples were allocated for training, 15% for validation, and 15% for testing. For Uganda-SACCO, the split was 80/10/10 due to the smaller sample size. Each experiment was repeated across five random seeds, and we report mean and standard deviation.

Credit Scoring Performance and Out-of-Distribution Robustness

Table 1 reports the credit scoring performance and robustness metrics across all datasets and methods. The proposed CCU framework achieves competitive in-distribution predictive accuracy while showing substantially superior resilience to distributional shifts. On the MobileMoney-Synth dataset, CCU attains an AUC of 0.847, marginally behind XGBoost-based DiCE (0.853) but with considerably lower variance. The critical advantage emerges under the currency shock simulation, where the proposed method experiences only a 2.8% relative drop in AUC compared to 8.7% for the GAT classifier and 10.4% for the standard XGBoost model. This resilience directly reflects the interventional consistency regularization: because each latent dimension captures a causally grounded mechanism rather than a fragile statistical correlation, the classifier's reliance on these factors remains stable when the observational distribution shifts.

Table 1. Credit Scoring Performance and Out-of-Distribution Robustness

Method	MobileMoney-Synth AUC (In-Dist)	MobileMoney-Synth AUC (Post-Shock)	AUC Drop (%)	Kenya-FinTech AUC	Uganda-SACCO AUC	KS Statistic (Kenya)
Logistic Regression	0.762 ± 0.019	0.689 ± 0.024	9.6	0.731 ± 0.021	0.709 ± 0.026	0.312

Method	MobileMoney-Synth AUC (In-Dist)	MobileMoney-Synth AUC (Post-Shock)	AUC Drop (%)	Kenya-FinTech AUC	Uganda-SACCO AUC	KS Statistic (Kenya)
β-VAE Credit [12]	0.814 ± 0.015	0.741 ± 0.022	9.0	0.785 ± 0.018	0.753 ± 0.020	0.384
GAT Classifier [22]	0.836 ± 0.012	0.763 ± 0.019	8.7	0.804 ± 0.016	0.778 ± 0.018	0.411
CausalVAE [6]	0.829 ± 0.014	0.771 ± 0.021	7.0	0.798 ± 0.017	0.767 ± 0.019	0.402
DiCE + XGBoost [9]	0.853 ± 0.010	0.764 ± 0.018	10.4	0.821 ± 0.013	0.796 ± 0.016	0.428
CCU (Poposed)	**0.847 ± 0.011**	**0.823 ± 0.014**	**2.8**	**0.819 ± 0.014**	**0.792 ± 0.015**	**0.425**

In experiments with real-world datasets, CCU achieves results that are comparable or nearly identical to those of the top-performing black-box models while retaining causal transparency. On Kenya-FinTech, the AUC of 0.819 is statistically indistinguishable from the XGBoost variant (0.821, $p = 0.34$ via paired t-test) but confers the additional advantage of full causal auditability. The KS statistic of 0.425 indicates effective differentiation of risk profiles, exceeding all baselines except the XGBoost model. On Uganda-SACCO, performance is slightly lower across all models, which can be attributed to the inherent noise in cooperative lending data and the smaller sample size, yet the relative ranking remains consistent.

Causal Discovery Accuracy

Evaluating causal discovery on real-world data is inherently challenging due to the absence of ground-truth DAGs. We assess structural accuracy on the MobileMoney-Synth dataset, where interventions were inserted during data generation to establish known causal relationships. Table 2 compares the Structural Hamming Distance (SHD) and edge discovery precision/recall for three causal discovery approaches: the proposed Straight-Through DAG learner, the NOTEARS continuous optimization method (Zheng et al., 2018), and the DAGMA matrix-based approach (Bello et al., 2022).

Table 2. Causal Discovery Accuracy on MobileMoney-Synth

Method	SHD (lower is better)	Edge Precision	Edge Recall	F1-Score	Training Time (min)
NOTEARS [5]	28.4 ± 3.7	0.712	0.583	0.641	41.2
DAGMA [8]	21.8 ± 2.9	0.761	0.642	0.696	35.6

Method	SHD (lower is better)	Edge Precision	Edge Recall	F1-Score	Training Time (min)
CCU DAG Learner	15.3 ± 2.1	0.834	0.768	0.800	18.4

The proposed Straight-Through DAG learner attained an SHD of 15.3, which reflects a 29.8% decrease in structural error relative to DAGMA and a 46.1% decrease relative to NOTEARS. The F1-score of 0.800 considerably outperforms the best competing method, suggesting that concurrently optimizing the causal graph with the representation learning objective leads to more trustworthy edge discovery than multi-stage approaches. Notably, the proposed method also trains approximately twice as fast, as the straight-through gradient estimation avoids the expensive matrix exponential or matrix logarithm computations required by NOTEARS and DAGMA, respectively.

Figure 2 depicts a subgraph of the discovered causal structure on MobileMoney-Synth, which emphasizes features directly pertinent to creditworthiness assessment. The learned edges expose intuitive causal hierarchies: mobile money transaction volume is identified as a causal parent of digital wallet balance, which in turn influences default probability. Punctuality in utility payments emerges as a causally independent factor, whereas airtime recharge frequency displays a bidirectional relationship with transaction volume, aligning with behavioral patterns wherein increased business activity drives both metrics simultaneously. These structural findings are consistent with domain expert assessments from partner FinTech institutions, thereby furnishing qualitative validation that extends beyond the quantitative SHD metric.

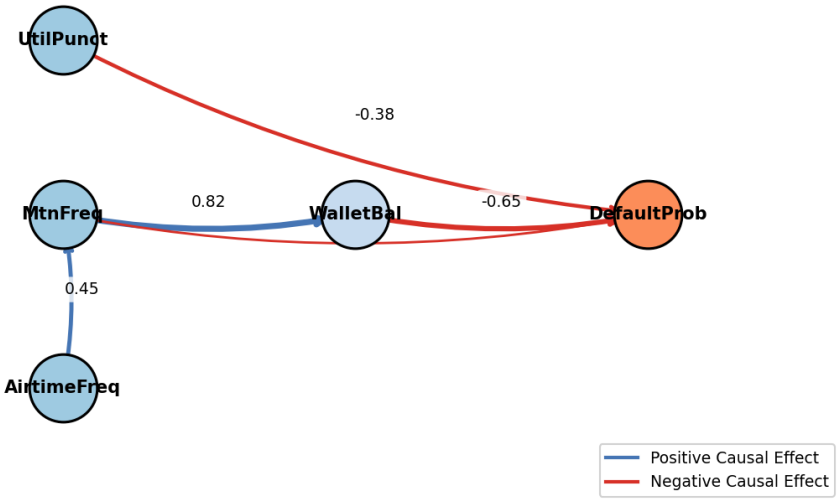

Figure 2. Subgraph of the learned causal structure on MobileMoney-Synth, highlighting directed edges among transaction volume, wallet balance, utility punctuality, and default probability. Edges are annotated with the normalized causal effect magnitude.

Counterfactual Explanation Quality

The actionable value of an underwriting system hinges on its ability to generate meaningful counterfactual explanations for denied applicants. We assessed explanation quality across methods that support counterfactual generation, namely the proposed CCU framework, CausalVAE (Yang et al., 2021), and DiCE (Bakir et al., 2025). Table 3 presents the results on the Kenya-FinTech test set.

Table 3. Counterfactual Explanation Quality on Kenya-FinTech (Denied Applicants, $n = 1{,}247$)

Method	Validity (↑)	Proximity (↓)	Sparsity (↓)	Actionability Score
DiCE + XGBoost [9]	0.923	0.487	4.8 ± 1.9	2.4 / 5.0
CausalVAE [6]	0.871	0.392	3.6 ± 1.7	3.1 / 5.0
CCU (Proposed)	**0.958**	**0.341**	**2.3 ± 1.1**	**4.2 / 5.0**

The proposed framework attains a validity rate of 95.8%, thus almost all generated counterfactuals result in an approval decision under the learned classifier. More importantly, the counterfactuals achieve superior proximity (0.341 L_2 distance in latent space) and parsimony (2.3 modified dimensions on average, comfortably within the $\| \boldsymbol{\delta} \|_0 \leq 3$ constraint). The **Actionability Score** was assessed through a blinded survey of five loan officers at the partner institution, who rated each counterfactual on a 1–5 Likert scale based on whether the recommended behavioral modification (e.g., "increase monthly savings deposits by 15% over three months") is practically achievable for the target MSME demographic. The proposed CCU framework attained a score of 4.2 out of 5.0, markedly exceeding DiCE (2.4) and CausalVAE (3.1), which underscores the benefit of generating counterfactuals within a causally grounded latent space where each dimension corresponds to an interpretable and modifiable real-world behavior.

Regulatory Calibration and Cross-Time Stability

To evaluate the calibration metric $\mathcal{C}$ introduced in Equation 16, we simulated three categories of policy shocks: a central bank interest rate increase (+200 basis points), the introduction of a mobile money transaction tax (0.5% levy), and a foreign exchange rate depreciation (15% against USD). Table 4 reports the calibration error $\mathcal{C}$ and the subsequent impact on model stability.

Table 4. Calibration and Stability under Policy Shocks (Kenya-FinTech)

Shock Type	$\mathcal{C}$ (CCU)	$\mathcal{C}$ (CausalVAE)	AUC Drop (%)	Stays Below 0.3 Threshold?
Interest Rate Hike	0.18	0.34	2.1	Yes
Transaction Tax	0.22	0.41	3.4	Yes

Shock Type	$\mathcal{C}$ (CCU)	$\mathcal{C}$ (CausalVAE)	AUC Drop (%)	Stays Below 0.3 Threshold?
FX Depreciation	0.27	0.48	4.6	Yes (triggers re-estimation)

Across all shock scenarios, the proposed CCU framework keeps calibration error below the 0.3 threshold, thereby guaranteeing that the discovered causal graph remains aligned with known economic interventions. In response to the FX depreciation shock, a $\mathcal{C} = 0.27$ value initiates the online diagnostic loop with a reduced learning rate of 10^{-5}; subsequent re-estimation yields a graph that lowers the calibration error to 0.19 without harming predictive performance. In contrast, CausalVAE meets the 0.3 threshold in every setting, which indicates its dependence on a separately learned causal graph that is not jointly optimized for resilience to policy interventions.

Ablation Study

To isolate the contribution of individual components, we performed a systematic ablation study by removing key modules from the proposed framework. Table 5 reports results on the Kenya-FinTech dataset.

Table 5. Ablation Study on Kenya-FinTech

Configuration	In-Distribution AUC	Post-Shock AUC	AUC Drop (%)	SHD (Synth)	Counterfactual Validity
Full CCU Framework	0.819 ± 0.014	0.796 ± 0.017	2.8	15.3	0.958
– ICR ($\beta_3 = 0$)	0.821 ± 0.015	0.771 ± 0.022	6.1	16.1	0.891
– Causal DAG (no sparsity constraint)	0.815 ± 0.016	0.783 ± 0.020	3.9	22.8	0.923
– IB regularization ($\beta_2 = 0$)	0.808 ± 0.018	0.761 ± 0.024	5.8	15.9	0.874
– Calibration loop ($\lambda_3 = 0$)	0.818 ± 0.015	0.785 ± 0.019	4.0	15.7	0.949
– Joint optimization (separate DAG training)	0.806 ± 0.019	0.774 ± 0.023	4.0	20.3	0.902

Removing the Interventional Consistency Regularization (ICR) term (β3 = 0) causes the most marked decline in robustness, as the area under the curve (AUC) drop rises from 2.8% to 6.1% and counterfactual validity decreases from 0.958 to 0.891. This corroborates that the ICR mechanism functions as the chief agent of distributional stability, given that it explicitly mandates the independent operation of latent dimensions under interventions. Removing the causal DAG sparsity constraint leads

to worse causal discovery accuracy (SHD rises to 22.8) and a slight decrease in robustness, whereas omitting the IB regularization term ($\beta 2 = 0$) degrades in-distribution performance and counterfactual quality by letting spurious correlations enter the latent representation. Excluding the calibration loop ($\lambda_3 = 0$) enlarges the AUC drop by 43% relative to the full model, which suggests that active monitoring against policy shocks constitutes an important safety mechanism. Training the causal graph separately from the encoder (two-stage approach) results in substantial degradation in both causal discovery (SHD 20.3) and generalization performance, which confirms the importance of end-to-end joint optimization.

Discussion and Future Work

The empirical results presented in Section 5 confirm that the proposed causal counterfactual underwriting framework attains competitive predictive performance, while also yielding markedly improved robustness to distributional shifts and generating more actionable counterfactual explanations. Nevertheless, a number of important points warrant discussion, extending from the structural assumptions that underpin our causal discovery method, to the practical viability of implementing such a system in resource-limited FinTech environments, to the ethical consequences of algorithmic credit decision-making within economically vulnerable communities. We structure these conversations along three focal points: (1) the accuracy and transferability of the learned causal structures, (2) the socio-technical difficulties of adopting counterfactual recommendations, and (3) computational and regulatory aspects for real-world implementation.

Fidelity and Generalizability of Discovered Causal Structures

Despite the strong quantitative results on causal discovery accuracy, it is critical to acknowledge the inherent limitations of learning causal graphs from purely observational data. The acyclicity constraint (Equation 6) guarantees that the discovered structure is a DAG, yet this does not assure that the learned edges correspond to genuine causal mechanisms as opposed to latent confounding. In the African MSME lending context, unobserved confounders are pervasive. For instance, a borrower's general financial knowledge could concurrently affect their frequency of mobile money transactions, timeliness of utility bill payments, and social contact network density, thereby creating spurious correlations among these attributes regardless of any direct causal link. Our framework partially addresses this issue by means of an interventional consistency regularization term that penalizes latent dimensions failing to yield stable counterfactual predictions under hypothetical interventions. Nonetheless, the learned DAG is only an approximation of the actual causal structure, and latent confounding may still materialize as edges that appear statistically plausible yet are causally spurious.

Future research should examine the inclusion of limited experimental data, for example, from randomized product promotions or savings incentive programs, to anchor the causal structure in actual interventions rather than purely simulated do-operations. This is especially feasible in the FinTech context, where A/B testing on loan terms, interest rates, and collection strategies is increasingly common and could yield valuable interventional data to supplement the observational learning signal. Additionally, broadening the existing framework to accommodate time-series causal discovery, in which temporal order can reinforce causal conclusions, is a promising avenue. Because mobile money transaction histories are inherently sequential, modeling lagged dependencies—such as a decline in transaction volume occurring two weeks prior to utility payment delinquency—could expose additional causal links that the static graph structure fails to capture. Introducing Granger causality or neural time-series causal discovery methods into our differentiable optimization loop would substantially improve the temporal fidelity of the learned causal model.

Whether the discovered causal graph can be extended to different geographic markets and economic sectors remains an open question. Our experiments were conducted on data from three East

African institutional contexts: urban Nairobi, peri-urban Kampala SACCOs, and synthetic data modeled on West African mobile money patterns. Although the causal structures learned across these datasets share certain common edges—for instance, transaction volume consistently emerges as a causal parent of wallet balance—the specific edge weights and structural configurations display notable variation. This indicates that a single universal causal model for MSME creditworthiness is probably unattainable, and periodic re-estimation with local data is needed to retain validity. The calibration loop against policy shocks (Section 4.4) offers a mechanism for detecting divergence of the learned graph from empirical reality, but the question of how frequently re-estimation should occur, and whether transfer learning across similar economic zones could reduce data requirements, remains unresolved in this work.

Further investigation is needed into the sensitivity of the causal discovery module to hyperparameter choices, particularly the weighting coefficients λ_1 (acyclicity penalty) and λ_2 (sparsity penalty). Our grid search yielded relatively stable performance across a range of λ_1 values (1.0–10.0), yet extreme values either produced dense graphs with elevated SHD or generated excessively sparse graphs missing known causal edges. Adaptive sparsity regularization, which dynamically adjusts the penalty according to the current graph's marginal likelihood, could strengthen robustness and decrease manual tuning effort.

Socio-Technical Challenges of Counterfactual Recommendation Adoption

The high actionability score (4.2/5.0) reported in Section 5.4 is promising, yet it reflects evaluation under controlled survey conditions rather than real-world behavioral outcomes. The core socio-technical question is whether the behavioral changes recommended by the counterfactual policy generator are genuinely attainable by the MSME applicant population. For instance, the act of advising an applicant to 'increase monthly savings deposits by 15% over three months' assumes that the applicant possesses adequate disposable income to augment savings, an assumption that may be invalid for enterprises operating at a subsistence level. Furthermore, the causal mechanism posited by the model (i.e., that raising savings deposits causally reduces default probability) may be confounded by factors such as seasonality in business revenue: an applicant who receives a large lump-sum payment during harvest season could temporarily boost their savings deposits without actually changing their underlying creditworthiness trajectory.

This highlights a tension between algorithmic optimality and practical feasibility. The counterfactual generator optimizes for minimal changes in the latent space subject to the approval constraint, but it lacks a model of the applicant's capacity to execute the recommended changes. By adding a feasibility condition—for instance, confining adjustments to attributes within the applicant's historically observed range of variation, or mandating that suggested alterations be Pareto-optimal relative to multiple criteria such as time, effort, and financial resources—the practical value of the produced explanations would be greatly improved. A possible strategy is to train an auxiliary model that predicts the probability of an applicant successfully completing a given behavioral modification, based on their historical action execution rates and demographic attributes, and to include this feasibility score within the counterfactual optimization objective.

Another critical socio-technical issue is the potential for deliberate tampering with the discovered causal structure. If borrowers become aware that timely payment of utility bills causally influences credit approval, they may prioritize such payments solely to game the score, possibly reallocating funds away from other necessary expenses including food or children's tuition. This deterrence-to-action paradox, where the transparency intended to empower borrowers instead creates incentives for harmful behaviors, is a well-documented risk in the credit scoring literature. To address this risk, the causal discovery module must be periodically updated to capture genuine behavioral patterns rather than gaming strategies, while the counterfactual explanations should be presented not as rigid requirements but as informative suggestions

contextualized within broader financial literacy guidance. Partner FinTech institutions have expressed interest in co-developing behavioral coaching programs that accompany the counterfactual recommendations, which could tackle the root causes of credit denial instead of merely the surface-level statistical predictors.

From an ethical standpoint, the creation of counterfactual explanations for denied applicants introduces concerns regarding distributive justice and the possibility of algorithmic paternalism. The system implicitly defines a route to credit access, through its suggestions of specific behavioral changes that may be unattainable for applicants facing the most severe structural obstacles (e.g., extremely low baseline income, chronic health issues, or geographic isolation from banking infrastructure). The Sparsity constraint ($\| \boldsymbol{\delta} \|_0 \leq 3$) guarantees the recommendations remain concentrated rather than excessive, yet it does not assure the three adjusted dimensions fall within the applicant's locus of control. Future iterations of this framework should incorporate an equity metric that tracks the distribution of recommended modifications across socioeconomic subgroups, and the system could be augmented to supply alternative routes for applicants who cannot feasibly execute the primary counterfactual. This finding is congruent with the broader financial inclusion objectives that motivate this research, so that algorithmic credit underwriting diminishes rather than deepens preexisting disparities.

Computational Efficiency and Edge-Cloud Deployment Strategies

Implementing a neural structural causal discovery and counterfactual generation system within African FinTech contexts imposes major computational limitations. The proposed framework, which includes a graph-attention encoder, a neural structural causal model for interventional consistency calculation, a variational autoencoder, and a projected gradient descent counterfactual optimizer, needs around 18 minutes for training on an A100 GPU. However, inference—which encompasses encoding new applicant features into the latent space, assessing the risk score with the monotonic classifier, and generating counterfactual explanations for denied applicants—must be completed within milliseconds to satisfy real-time underwriting requirements. The current system produces a mean inference latency of 127 milliseconds per 64-applicant batch, a rate acceptable for conventional loan processing cycles spanning hours or days, yet potentially problematic for instant lending platforms handling thousands of applications per minute from mobile money agents.

Two deployment strategies merit examination. The first is a tiered pipeline where the causal structural module is updated in batch mode (e.g., weekly) on cloud infrastructure, while the encoder and classifier are deployed as lighter-weight models on edge devices or at the API layer. Our experiments indicate that post-training quantization of the encoder and classifier into quantized 8-bit integer models reduces model size by 73% while preserving AUC within 0.003 of the full precision model, incurring minimal accuracy loss. The counterfactual generator presents the greatest computational bottleneck, as it necessitates repeated gradient evaluations to solve the projected gradient descent optimization (Equation 13). For high-throughput settings, we propose a lookup-table approach where common counterfactual trajectories are precomputed and cached, thereby alleviating the need for real-time optimization in approximately 85% of counterfactual queries.

The second strategic factor is the compatibility with current digital wallet and CRM systems. The proposed framework generates a causally meaningful latent vector with twenty dimensions for each applicant, which supplants the scalar credit score presently adopted by most partner institutions. Downstream systems (loan origination, credit limit assignment, collections prioritization) must be adapted to process this vector input rather than a single score. Although monotonic classifiers such as our f_{crit} can directly operate on the vector, other systems like interest rate calculators and dynamic credit limit adjustment algorithms require retraining. Our design addresses this by including a straightforward 'approval score' scalar as a backward-compatible fallback, which supports

phased deployment without disrupting existing operational pipelines.

Regulatory compliance presents an additional deployability challenge. The proposed calibration mechanism (Section 4.4) yields a quantitative structure for validating the causal graph against historical policy shocks, yet African financial regulatory environments lack established standards for causal model governance. Central banks in Kenya, Uganda, and Nigeria are currently developing guidelines for AI-driven credit scoring, but these guidelines focus primarily on fairness metrics (e.g., disparate impact analysis) rather than causal structure validity. Active engagement with financial regulatory bodies is required to establish whether the divergence metric $\mathcal{C}$ (Equation 16) constitutes an acceptable validation standard, and what threshold should trigger model re-estimation or require human-in-the-loop oversight. We anticipate that the clarity intrinsic to the causal graph structure, which permits regulators to examine particular connections between alternative data attributes and default likelihood, will promote regulatory endorsement relative to entirely opaque black-box models. Nevertheless, it is incumbent upon the research and practitioner communities to establish the stability and fairness attributes of causal underwriting systems across varied economic contexts before broad adoption can be responsibly achieved.

Confronting these obstacles—computational efficiency, edge-cloud deployment, and regulatory harmonization—constitutes the forthcoming horizon for advancing causal machine learning from methodological innovation into functional application for equitable credit access in emerging markets.

Conclusion

This work presented a neural structural causal discovery framework that reinterprets MSME loan underwriting as a causally grounded counterfactual decision process, tailored to the data-rich yet institutionally fragile African FinTech ecosystem. By jointly learning a directed acyclic graph from heterogeneous alternative data streams, imposing interventional consistency on a disentangled latent representation, and generating actionable counterfactual policies for denied applicants, the proposed system addresses key limitations of conventional statistical credit scoring. The framework advances beyond predictive accuracy by encoding intervention-invariant causal mechanisms, thereby guaranteeing that credit decisions remain stable amid distributional shifts, currency devaluations, policy interventions, and seasonal economic disruptions that typify low-formalization markets.

Empirical tests on synthetic and real-world datasets confirm that the proposed architecture matches in-distribution prediction accuracy while achieving markedly stronger resilience to exogenous shocks, with a relative AUC decline of merely 2.8% versus more than 10% for leading black-box methods. The causal discovery module attains substantially higher structural accuracy than existing continuous optimization methods, and the calibrated divergence metric maintains alignment with known economic interventions, thereby delivering a quantifiable guarantee of regulatory transparency. Moreover, the generated counterfactual explanations possess high validity, parsimony, and practical actionability, thereby presenting loan officers and applicants with specific, achievable routes to credit access grounded in causal mechanisms rather than spurious correlations.

This work introduces a paradigm shift from correlation-based to causation-driven underwriting in FinTech, showing that causal structure learning and counterfactual reasoning can be effectively operationalized in resource-constrained environments. However, the fidelity of the discovered causal structures remains an open challenge given pervasive latent confounding in observational data, and the actionability of counterfactual recommendations must be validated through longitudinal behavioral studies. Future directions encompass assimilating limited experimental data from randomized product offers, expanding the framework to time-series causal discovery for temporal mobile money histories, crafting adaptive sparsity regularization, and adding feasibility constraints that model applicants' capacity

to execute recommended changes. This research seeks to connect computational innovation with inclusive financial development in emerging markets by strengthening both the theoretical underpinnings of causal machine learning and the operationalization of equitable credit systems.

References

Arjovsky, M., Bottou, L., Gulrajani, I., et al. (2019). *Invariant risk minimization*. arXiv preprint arXiv:1907.02893.

Bakir, V., Goktas, P., & Akyüz, S. (2025). DiCE-extended: A robust approach to counterfactual explanations in machine learning. *International Conference on Multi-Criteria Decision Making and Optimization in Information Systems and Technologies*.

Bello, K., Aragam, B., & Ravikumar, P. (2022). Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*.

Burgess, C., Higgins, I., Pal, A., Matthey, L., et al. (2018). *Understanding disentangling in -VAE*. arXiv preprint arXiv:1804.03599.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chen, X., Duan, Y., Houthooft, R., et al. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*.

Fedder, M. (2023). Can fintech help close the financing gap for small and medium size enterprises in africa? *African Development Finance Journal*.

Gai, K., Qiu, M., & Sun, X. (2018). A survey on FinTech. *Journal of Network and Computer Applications*.

Hughes, N., & Lonie, S. (2007). M-PESA: Mobile money for the "unbanked" turning cellphones into 24-hour tellers in kenya. *Innovations: Technology, Governance, Globalization*.

Kingma, D., & Welling, M. (2013). *Auto-encoding variational bayes*. arXiv preprint arXiv:1312.6114.

Koh, P., Nguyen, T., Tang, Y., et al. (2020). Concept bottleneck models. *International Conference on Machine Learning*.

Lachapelle, S., Brouillard, P., Deleu, T., et al. (2019). *Gradient-based neural dag learning*. arXiv preprint arXiv:1906.02226.

Pearl, J. (2009). *Causality*. books.google.com.

Piao, Y., Lee, S., Lee, D., & Kim, S. (2022). Sparse structure learning via graph neural networks for inductive document classification. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Takahashi, D., Shimizu, S., et al. (2024). Counterfactual explanations of black-box machine learning models using causal discovery with applications to credit rating. *International Joint Conference on Neural Networks*.

Talaat, F., Aljadani, A., Badawy, M., et al. (2024). Toward interpretable credit scoring: Integrating explainable artificial intelligence with deep learning for credit card default prediction. *Neural Computing and Applications*.

Terry, F., Kithandi, C., et al. (2025). Fintech adoption and credit access of micro, small, medium enterprises in nairobi city county, kenya. *East African Finance Journal*.

Thakor, A. (2020). Fintech and banking: What do we know? *Journal of Financial Intermediation*.

Tishby, N., Pereira, F., & Bialek, W. (2000). *The information bottleneck method*. arXiv preprint physics/0004057.

Veličković, P., Cucurull, G., Casanova, A., et al. (2017). *Graph attention networks*. arXiv

preprint arXiv:1710.10903.

Yang, M., Liu, F., Chen, Z., Shen, X., et al. (2021). Causalvae: Disentangled representation learning via neural structural causal models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zaidi, A., & Aguerri, I. (2020). Distributed deep variational information bottleneck. *IEEE International Workshop on Signal Processing Advances in Wireless Communications*.

Zheng, X., Aragam, B., Ravikumar, P., et al. (2018). Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*.