Beyond Attribution: A Falsifiability Framework for Reliable Credit Risk Explanations

Anonymous Author(s)

***Abstract*—Machine learning explanations are often treated as credible outputs of model introspection, yet they may reflect statistical artefacts rather than genuine learned patterns. This study addresses a critical gap in credit-risk governance: the need for scientifically rigorous, testable explanations rather than post-hoc narratives. We propose a falsifiability framework that treats explanations as empirical hypotheses subject to systematic validation. The framework operationalises this through three epistemically distinct components: (1) dual-selector feature stabilisation combining nonlinear importance scoring with sparse linear regularisation; (2) reliability diagnostics including sanity ratio validation and signal-to-noise discrimination; and (3) a constrained generative AI module that grounds natural language explanations in validated SHapley Additive exPlanations (SHAP) evidence. By integrating explanation reliability assessment directly into model development—rather than treating it as an afterthought—we demonstrate that strong predictive performance and reliable explanations are orthogonal properties requiring independent scrutiny. Our experimental validation across four heterogeneous credit-risk datasets—German Credit, Bank Marketing, Credit Approval, and Statlog (Australian Credit)—spanning sample sizes from 300 to 45,211 observations shows that high-performing models can produce feature attributions with widely varying reliability, ranging from near-noise to strong signal, depending on dataset characteristics. This cross-dataset evidence establishes that the prediction–explanation decoupling is not an artefact of a single benchmark but a systematic phenomenon requiring dataset-aware reliability diagnostics. This work enables transparent, auditable explanations suitable for regulatory compliance and stakeholder trust, fundamentally advancing how financial institutions govern machine learning systems.**

***Index Terms*—Credit Risk, Explainability, Reliability, Falsifiability, SHAP**

I. Introduction

CREDIT-RISK assessment plays a central role in financial decision-making, shaping lending policies and capital allocation in regulated institutions [3], [6]. Contemporary machine-learning models have substantially improved predictive discrimination, yet their explanations remain epistemically untested. A widespread assumption persists in practice: strong predictive performance implies reliable model explanations. This assumption is false. A model's ability to discriminate borrowers by default risk does not guarantee that its feature attributions reflect genuine learned structure rather than statistical artefacts, sampling noise, or spurious correlations [14]. This disconnect between prediction and explanation constitutes a fundamental scientific problem: explanations are treated as credible outputs rather than as empirical hypotheses subject to rigorous validation.

The challenge is compounded by the proliferation of post-hoc explainability tools and generative AI. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) assign numerical attributions to features, addressing some interpretability concerns in high-accuracy ensemble models [9], [10]. However, growing evidence indicates that these attribution methods are highly sensitive to background distributions and sampling noise, frequently reflecting artefacts rather than genuine model structure [12], [14]. Recent applications of generative AI that translate attribution scores into natural-language explanations remain largely unconstrained, further amplifying risks of hallucination, narrative inflation, and spurious causal claims [14]. The result is a proliferation of confident-sounding explanations with no mechanism to assess their empirical grounding.

This fragmentation exposes a deeper methodological gap: explainability in credit-risk modelling is rarely treated as a falsifiable scientific process. Existing work predominantly frames explanations as descriptive summaries or visualisation artefacts rather than as testable hypotheses subject to systematic evaluation. No prior framework integrates predictive modelling, attribution stability assessment, and constrained generative AI reasoning while explicitly quantifying whether explanatory signals exceed noise baselines under empirical scrutiny [12], [14].

This paper addresses this gap by proposing a falsifiability framework that reframes explainability from narrative generation to empirical validation. The core insight is simple but consequential: treat explanations as claims subject to reliability testing. The framework introduces three epistemically distinct components: (1) dual-selector feature stabilisation that grounds attributions in both nonlinear interaction signals and sparse linear structure prior to model training; (2) reliability diagnostics based on sanity ratio validation and signal-to-noise discrimination, enabling explicit assessment of whether attribution signals are robust or spurious; and (3) a constrained generative AI module that produces human-readable explanations qualified by explicit uncertainty statements when reliability diagnostics indicate weak attribution evidence. This ensures stakeholders receive transparent assessments of explanatory confidence rather than unfounded certainty.

To stabilise the features used for explanation, the framework adopts the Feature Selector-classifier Optimisation Framework proposed by Zeng *et al.* [19]. Their dual-selector mechanism—combining nonlinear Random Forest importance with sparse L1-regularised logistic regression coefficients—establishes a

Manuscript received Month Day, Year; revised Month Day, Year.
Author information withheld for double-anonymous review.
Digital Object Identifier will be inserted by IEEE.

stable feature foundation prior to SHAP analysis. This hybrid design reduces estimator bias while preserving both interaction-aware and linear structural signals, creating the epistemic conditions necessary for trustworthy attribution.

To assess generalisability, the framework is evaluated across four heterogeneous credit-risk datasets—German Credit [21], Bank Marketing [22], Credit Approval [23], and Statlog (Australian Credit) [24]—spanning sample sizes from 300 to 45,211 observations, varying feature structures (numerical, categorical, mixed), and diverse class-imbalance profiles [3], [6]. A broad family of calibrated classification models is applied independently to each dataset, enabling systematic assessment of both predictive performance and the empirical reliability of model explanations across distinct data environments. The results reveal a structural decoupling: explanation quality varies independently of predictive accuracy, and even high-performing models can produce feature attributions with widely varying reliability depending on dataset characteristics. By providing explicit, quantifiable diagnostics for explanation reliability across multiple benchmarks, this work establishes a scientifically rigorous approach to explainability—one that treats it as a testable, falsifiable component of model validity subject to the same empirical standards as predictive performance. The implications extend to regulatory compliance: financial institutions can now ground model governance in transparent, auditable explanations rather than in subjective narratives.

II. Literature Review

Research on credit-risk modelling has evolved along two largely disconnected trajectories: optimisation of predictive algorithms and development of post-hoc interpretability methods. This fragmentation has created a critical blind spot: although robust benchmarks for predictive accuracy are well established, the scientific standards for assessing explanation reliability remain underdeveloped. Methodological fragmentation persists: studies emphasising predictive discrimination often sideline interpretability altogether, while explainability-focused work frequently frames explanations as descriptive narratives rather than as testable claims. No systematic approach yet integrates explanation reliability as a first-class validation concern alongside predictive performance. This review synthesises these strands to motivate the unified predictive–explanatory framework proposed in this study, which positions explanations as empirical hypotheses subject to falsifiable testing.

A. Predictive AI Research and Feature Optimisation

Early credit-risk models relied on classical statistical techniques such as logistic regression and linear discriminant analysis, valued for transparent coefficient structures [2]. However, these approaches struggle to capture nonlinear interactions and heterogeneous borrower behaviour. Comparative benchmarks, notably by Baesens *et al.* [3], consistently show that such linear assumptions underperform relative to flexible machine-learning models.

As a result, ensemble-based methods—including Random Forest, Gradient Boosting, XGBoost, and LightGBM—have become dominant in credit scoring, delivering substantial gains in discriminatory power (AUC) and separation efficiency (KS) [6], [5]. Although deep learning has been explored, evidence indicates that for modest tabular datasets such as German Credit, well-tuned tree ensembles and regularised linear models often achieve superior discriminatory power and calibration quality compared to more complex architectures [7], [4].

Feature stability has emerged as a critical yet underemphasised determinant of both predictive and explanatory robustness. This is consequential for interpretability: if feature rankings themselves are unstable across random seeds or background distributions, then SHAP attributions computed on those features inherit that instability, rendering explanations unreliable even if the model's predictive accuracy is high [13], [11]. Addressing this, Zeng *et al.* [19] proposed a Feature Selector-classifier Optimisation Framework that couples feature selection techniques with ensemble classifiers. Their dual-selector approach stabilises the feature foundation before model training, reducing estimator bias while preserving both nonlinear interaction signals and sparse linear structure. This principle is critical for explanation reliability: downstream explanations built on unstable features will themselves be unstable, regardless of model accuracy [13]. This study adopts this principle to ground downstream SHAP analysis in stable, validated predictive signals rather than in unstable single-estimator rankings.

Robustness is further shaped by the handling of class imbalance. Methods such as SMOTE can improve minority-class detection without degrading generalisation, provided they are applied strictly within stratified cross-validation to prevent information leakage [1], [8].

B. The Interpretability Gap and the Reliability Problem

Interpretability is both a regulatory and practical requirement in credit risk. Regulatory frameworks including Basel model risk management principles [16], SR 11-7 guidance [17], and EBA discussions on machine learning for internal ratings-based models [18] emphasise the need for transparent, auditable model explanations. While traditional models offered intrinsic interpretability [15], the opacity of modern ensemble methods has driven reliance on post-hoc attribution tools. However, this shift has created a subtle but consequential problem: interpretability (the ability to describe what a model does) has become conflated with reliability (the assurance that those descriptions are trustworthy). These are orthogonal properties.

LIME [9] and SHAP [10] have become standard approaches for explaining black-box models by assigning local feature attributions. These methods are commonly used to assess the economic plausibility of model drivers [20]. However, growing evidence reveals a fundamental problem: these attribution methods produce confident-sounding outputs regardless of whether the underlying signal is robust or noisy. Hassija *et al.* [14] demonstrate that attribution scores often conflate signal and noise, while Slack *et al.* [12] show that they are vulnerable

to adversarial manipulation, raising concerns for regulated deployment. Critically, no standard methodology distinguishes between attributions driven by genuine learned structure and those driven by statistical artefacts.

The emergence of generative AI to translate attribution scores into natural-language explanations has amplified this problem. These approaches typically lack epistemic constraints and remain susceptible to hallucination and narrative inflation [14]. The result is a proliferation of confident-sounding explanations with no mechanism to assess their empirical grounding. A model can achieve state-of-the-art predictive accuracy while producing feature attributions that are internally inconsistent or driven primarily by noise rather than signal—yet practitioners have few tools to detect this failure.

The field treats explainability as an interpretive challenge (how to describe predictions) rather than as a validation challenge (whether those descriptions are empirically grounded). Prior work addresses individual components—predictive optimisation [6], [19], attribution methods [9], [10], or generative AI explanation [14], [20]—but no systematic approach integrates these within a unified framework that subjects explanations to rejectable reliability diagnostics. This study addresses this gap.

III. Methodology

This study adopts a unified predictive–explanatory architecture grounded in falsifiability principles to benchmark credit-risk models while explicitly evaluating the reliability of their explanations. The framework integrates a calibrated predictive pipeline across multiple algorithmic families with a dual-selector feature stabilisation layer and a constrained generative AI explanation module. Critically, the methodology operationalises falsifiability by embedding reliability diagnostics as quantifiable rejection conditions: explanations are always generated, but confident claims are suppressed when signal-to-noise discrimination indicates weak attribution evidence, and generated narratives are accompanied by explicit uncertainty qualifications. This ensures that predictive performance, attribution stability, and explanatory uncertainty are assessed within a single coherent workflow oriented toward empirical validation rather than narrative confidence.

A. Architectural Principles: Conceptual Foundation

The three-layer architecture shown in Fig. 1 is a *non-procedural* representation of task decomposition and information dependencies. The three layers comprise: (1) feature stabilisation via dual-selector screening, (2) model training and selection across calibrated ensemble families, and (3) explanation generation with reliability diagnostics. The architecture specifies what transformations are logically necessary and how information must flow between stages, but does not impose ordering constraints, quantitative selection criteria, or conditional rejection logic. Any valid instantiation must respect the architectural constraints (e.g., feature stabilisation logically precedes model training), but the specific operational realisation—including hyperparameter selection, decision thresholds, and failure handling—is determined by the algorithm.

TABLE I: Dataset Characteristics

Dataset	Records	Feat.	Type	Good	Bad
German Credit [21]	1,000	20	7N, 13C	700	300
Bank Marketing [22]	45,211	16	Mixed	39,922	5,289
Credit Approval [23]	300	15	Cat/Int/Real	211	89
Statlog (Australian) [24]	690	14	6N, 8C	383	307

B. Data and Preprocessing

To assess the generalisability of the falsifiability framework across diverse data environments, four publicly available credit-risk datasets are selected from the UCI Machine Learning Repository. These datasets are chosen to span a wide range of sample sizes, feature structures, class-imbalance profiles, and domain contexts, ensuring that the framework's conclusions are not artefacts of a single benchmark. Table I summarises the key properties of each dataset.

The **German Credit** dataset [21], [3] is a widely used benchmark in credit-risk research comprising 1,000 observations and 20 attributes (7 numerical, 13 categorical), with a 70:30 good-to-bad class split. Its moderate size and complexity make it well suited for thorough reliability assessment without computational barriers to exhaustive cross-validation and sanity checking.

The **Bank Marketing** dataset [22] comprises 45,211 observations and 16 attributes derived from direct marketing campaigns of a Portuguese banking institution. The target variable indicates whether the client subscribed to a term deposit. Its substantially larger sample size and severe class imbalance (88.3% negative) test the framework's scalability and robustness to highly skewed distributions.

The **Credit Approval** dataset [23] contains 300 observations (selected subset) with 15 attributes of mixed types (categorical, integer, and real-valued). Its small sample size creates a challenging environment for both predictive modelling and reliability assessment, testing whether the framework's diagnostics remain informative under limited-data conditions.

The **Statlog (Australian Credit)** dataset [24] comprises 690 observations with 14 attributes (6 continuous, 8 categorical) and a near-balanced class distribution (55.5% good, 44.5% bad). This near-balance provides a contrasting data environment to the imbalanced datasets above, enabling assessment of whether class distribution affects explanation reliability.

Together, these four datasets create a comprehensive experimental testbed spanning sample sizes from 300 to 45,211, feature counts from 14 to 20, and class-imbalance ratios from near-balanced to highly skewed. This diversity enables systematic investigation of whether the prediction–explanation decoupling observed in prior single-dataset studies is a general phenomenon or an artefact of specific data characteristics.

To ensure robust model estimation and prevent confounding of predictive and explanatory reliability, all datasets undergo a standardised preprocessing sequence grounded in a principle of minimal assumptions: attributes with more than 90% missing values are removed (eliminating spurious correlations from sparse data), while remaining numerical and categorical missing values are imputed using median and mode strategies, respectively. Categorical variables are transformed using one-

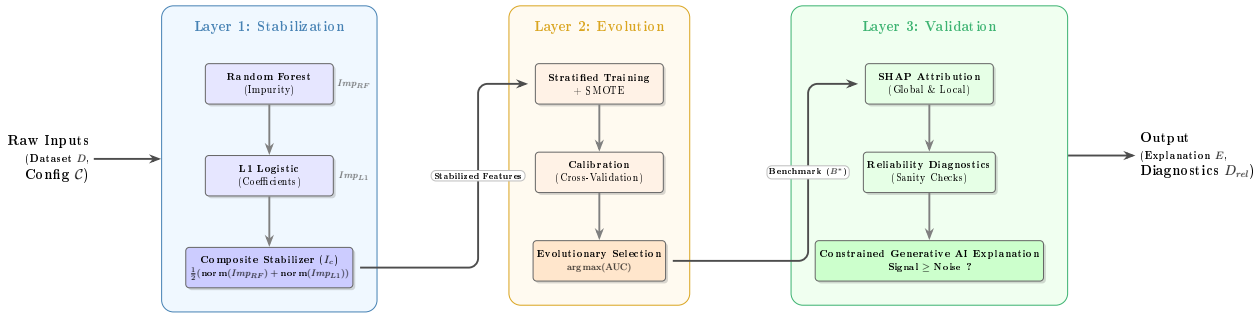


Fig. 1: Conceptual Architecture: Task Decomposition and Information Constraints.

hot encoding, and numerical features are standardised to zero mean and unit variance. This conservative preprocessing preserves data structure without introducing artificial associations that could confound downstream SHAP analysis. The identical preprocessing pipeline is applied to each dataset to ensure that cross-dataset comparisons reflect genuine differences in data characteristics rather than methodological artefacts.

Class imbalance is addressed using the Synthetic Minority Over-sampling Technique (SMOTE), applied separately within each training fold during stratified cross-validation for all four datasets. This placement is critical: applying SMOTE before splitting would contaminate test-set evaluations with synthetic data, undermining the empirical validation of both predictive performance and explanation reliability. By applying SMOTE only within training folds, we ensure that performance estimates and reliability diagnostics reflect genuine generalisation rather than artefacts of resampling [1], [8]. The effect of SMOTE is expected to vary across datasets given their different imbalance profiles, particularly for the severely imbalanced Bank Marketing dataset.

C. Predictive Modelling Framework

The unit of analysis is the model family, interpreted as a functional constructor that defines a class of predictive algorithms sharing a common architectural principle. Four primary model families are evaluated: Linear Models, Boosting, Bagging, and Instance-Based Learners. Hyperparameter choices (e.g., ensemble size, learning rate, regularisation structure, distance weighting) represent implementation variants of the same constructor and are not treated as independent hypotheses, but rather as necessary operational specifications that instantiate the family's functional definition.

To establish a comprehensive baseline that adequately exercises each family's representational capacity, 75 calibrated model configurations are systematically evaluated across these families, with the configuration space summarised in Table II. The same model registry is applied independently to each of the four datasets, ensuring that cross-dataset comparisons reflect genuine differences in data characteristics rather than methodological inconsistencies. This sampling strategy ensures that each model family is assessed fairly by exploring its operationally relevant hyperparameter ranges. All models are trained within a stratified cross-validation framework and calibrated using CalibratedClassifierCV to ensure that predicted scores correspond to well-formed probability estimates, a prerequisite for meaningful risk ranking and expected-loss interpretation.

TABLE II: Model families as functional constructors and their implementation variant dimensions.

Model Family	Constructor Implementation Dimensions	Configs
Linear	Solver and regularisation in logistic regression (lbfgs, saga, newton-cg; L1, L2, ElasticNet)	6
Boosting	Ensemble size, learning rate for AdaBoost; ensemble size for stochastic gradient boosting	28
Bagging	Ensemble size for bagged trees and neural nets; ensemble size and feature subsampling for random forests	36
Instance	Neighbourhood size and distance weighting in k-NN, with CV tuning	5
Total	**Combined instantiations across constructors**	**75**

1) Unit of Analysis: Model Families as Functional Constructors: A critical conceptual distinction underlies the experimental design and the falsifiability framework: the unit of analysis is not the individual trained model with specific hyperparameter values, but the model family itself, understood as a functional constructor that encodes an algorithmic principle. This distinction matters for reliability assessment. Hyperparameter choices (ensemble size, learning rate, regularisation magnitude) do not constitute independent competing hypotheses; instead, they represent operational implementation variants necessary to instantiate the constructor's functional definition. For example, varying ensemble size in a Bagging constructor does not generate a distinct hypothesis but rather explores how the bagging principle scales across different operational regimes. Similarly, varying regularisation in logistic regression does not test whether regularisation is correct—it tests how much regularisation best instantiates the linear-classification principle for this particular dataset.

This interpretation ensures that comparative assessment is conducted at the appropriate level of abstraction: families are compared as distinct algorithmic approaches, while within-

family variation documents the family's effective operating envelope and operational constraints. The selection of the best instantiation within each family identifies the most effective realisation of that family's core architectural principle. Crucially, this design prevents confusing poor hyperparameter choices with fundamental algorithmic failure: when an explanation is flagged as unreliable, we ask whether the reliability problem is intrinsic to the family's principle or contingent on suboptimal instantiation.

D. Explainability and Evaluation Architecture

The framework extends beyond predictive benchmarking by embedding explanation reliability assessment directly into the evaluation pipeline as a testable, rejectable component. Rather than treating feature attributions as self-validating artefacts, explanations are interpreted as empirical claims whose credibility depends on the empirically measured stability and strength of the underlying signal. This operationalises falsifiability: explanations are always generated but are explicitly subjected to signal-quality tests, with rejection conditions specified in advance (Sanity Ratio thresholds, noise discrimination criteria). When reliability diagnostics indicate robust signal, explanations are accompanied by confidence metrics; when diagnostics indicate weak signal, explanations are qualified with explicit uncertainty statements, creating a transparent audit trail of explanatory confidence.

1) Sanity Ratio: Formal Definition and Reliability Criterion: The Sanity Ratio ρ is a model randomisation diagnostic that quantifies whether SHAP attributions reflect genuine learned structure or are indistinguishable from noise. It is defined as:

$$\rho = \frac{\bar{S}_{real}}{\bar{S}_{rand}}, \qquad \bar{S} = \frac{1}{n}\sum_{i=1}^{n} |\phi_i| \tag{1}$$

where $\bar{S}_{real}$ denotes the mean absolute SHAP value computed from a model trained on true labels, $\bar{S}_{rand}$ denotes the mean absolute SHAP value from an architecturally identical model trained on randomly permuted labels $\tilde{y} = \sigma(y)$, ϕ_i are the individual SHAP attributions, and n is the number of observations in the evaluation set. The permutation σ destroys all feature–label associations while preserving marginal feature distributions, ensuring that $\bar{S}_{rand}$ provides a noise baseline under the null hypothesis of no learned signal.

If the model has learned genuine predictive structure, its SHAP attributions should be considerably larger and more concentrated than those of a model trained on random noise. Consequently, $\rho \gg 1$ indicates reliable explanations, while $\rho \approx 1$ indicates that real attributions are indistinguishable from the noise baseline—a critical reliability failure. Values $\rho < 1$ indicate that the real model's attributions are actually weaker than the random baseline, representing the most severe reliability failure.

To translate ρ into a normalised reliability weight, the signal weight $W_{signal} \in [0, 1]$ is defined using logarithmic scaling:

$$W_{signal} = \begin{cases} 0, & \rho \leq 1 \\ \min\left(\dfrac{\ln \rho}{\ln 3},\ 1\right), & \rho > 1 \end{cases} \tag{2}$$

Logarithmic scaling is adopted because it enforces an explicit noise threshold at $\rho = 1.0$ (where $\ln(1) = 0$, yielding $W_{signal} = 0$ regardless of attribution magnitudes), correctly penalises ratios only marginally above 1, and saturates at $W_{signal} = 1$ for $\rho \geq 3$, reflecting diminishing marginal returns on signal quality. To illustrate, a value of $\rho = 1.005$ would yield $W_{signal} = \ln(1.005)/\ln(3) \approx 0.005$, confirming attributions only marginally above the noise floor and triggering the uncertainty-qualified explanation pathway in Phase 4 of Algorithm 1.

2) Feature Attribution and Generative AI Explanation: To mitigate instability associated with single-method feature selection and prevent attribution uncertainty from being misattributed to true signal loss, a dual-selector mechanism is employed. By combining impurity-based Random Forest importance (which captures nonlinear interactions and heterogeneous effects) with coefficient-based L1-regularised logistic regression importance (which enforces sparse, interpretable linear structure), the framework preserves both interaction-aware and linear structural signals. This dual approach grounds subsequent SHAP analysis in stable feature foundations rather than in unstable single-estimator rankings. SHAP values are then computed on this stabilised feature set and passed to a generative AI module that translates quantitative attributions into human-readable narratives. Critically, the generative AI component is constrained by the reliability diagnostics: confident explanations are produced only when sanity-ratio and signal-to-noise tests pass, preventing the LLM from generating confident narratives when underlying attribution signals are weak.

The architecture is instantiated through a concrete operational procedure that specifies the exact sequencing, selection criteria, and failure conditions under which a valid model and explanation are produced. This procedure is summarised in Algorithm 1. Unlike the architecture, the workflow is falsifiable: it specifies quantitative selection criteria (AUC maximisation), computational order (feature stabilisation before model training), and explicit rejection conditions (explanation suppression when Sanity Ratio indicates weak signal).

E. Evaluation Metrics

Model evaluation in credit-risk modelling requires multidimensional assessment reflecting both regulatory requirements and practical decision-making demands. We employ a suite of complementary metrics that jointly capture discrimination, calibration, and cost-sensitive performance.

AUC serves as the primary model selection metric, chosen not for its optimality but for its falsifiability properties. AUC measures discriminatory power—the model's ability to rank-order observations by risk—independent of decision thresholds, providing a metric that is robust to class imbalance and threshold selection artefacts. This threshold-independence is critical for fair comparison across heterogeneous model families and enables reproducible selection logic.

However, AUC alone does not characterise explanatory reliability. The Brier Score complements AUC by quantifying probability calibration quality, measuring whether predicted

Algorithm 1: Operational Instantiation: Unified Predictive and Explanatory Workflow

Input: Dataset collection $\mathcal{D} = \{D_1, \ldots, D_K\}$, Model Registry $\mathcal{M}$, Configuration Set $\mathcal{C}$
Output: Per-dataset benchmark models $\{B_k^*\}$, Explanations $\{E_k\}$, Diagnostics $\{D_{rel,k}\}$, Cross-dataset comparison $\mathcal{R}$
for *each dataset* $D_k \in \mathcal{D}$ **do**
  Phase 1: Feature Screening (Dual-Selector)
  Train RF and L1-LR on D_k; compute composite score $I_c = \frac{1}{2}(\text{norm}(Imp_{RF}) + \text{norm}(Imp_{L1}))$; rank and select feature set.
  Phase 2: Model Training and Selection
  for *each family* $F \in \mathcal{M}$ **do**
    Stratified split; SMOTE within fold; train calibrated instantiations; evaluate AUC.
  $B_k^* = \arg\max_B(\text{AUC}(B))$
  Phase 3: Explainability and Reliability Assessment
  Compute SHAP values for B_k^*. Train shadow model on permuted labels $\tilde{y}$; compute $\rho_k = \bar{S}_{real}/\bar{S}_{rand}$ where $\bar{S} = \frac{1}{n}\sum|\phi_i|$
  Compute $W_{signal,k} = \min(\ln \rho_k / \ln 3,\ 1)$ for $\rho_k > 1$, else 0
  threshold $\theta_W = 0.10$
  Phase 4: Constrained Generative AI Explanation
  if $W_{signal,k} \geq \theta_W$ **then**
    Invoke LLM grounded in validated SHAP evidence
  else
    Invoke LLM with explicit uncertainty qualifications
Phase 5: Cross-Dataset Comparative Analysis
Aggregate $\mathcal{R} = \{(\rho_k, W_{signal,k}, B_k^*, \text{AUC}_k)\}_{k=1}^K$; assess consistency of prediction–explanation decoupling.
return $\{B_k^*, E_k, D_{rel,k}\}$, $\mathcal{R}$

default probabilities align with empirical frequencies across the full probability spectrum. Calibration and discrimination are orthogonal properties: a model can discriminate perfectly (rank order) while assigning poorly calibrated probabilities, or vice versa. By tracking both, we create the conditions for falsifying the claim that "high AUC implies reliable explanations." If a model achieves high AUC but poor Brier Score, we expect its feature attributions to exhibit anomalies under sanity checking.

All metrics are computed within stratified cross-validation to ensure that performance estimates reflect genuine generalisation rather than training-set artefacts. Secondary metrics provide additional diagnostic perspectives: the Kolmogorov-Smirnov (KS) statistic measures maximum separation between cumulative score distributions for default and non-default cases; the H-measure accounts for class imbalance and threshold selection effects [15]; and Recall quantifies the proportion of actual defaults correctly identified. The full metric suite creates a multi-dimensional performance space in which no single model dominates all dimensions, forcing practitioners to make trade-off choices and preventing false claims of universal optimality.

IV. Results

This section reports the empirical findings of the proposed predictive–explanatory framework applied independently to four credit-risk datasets (Table I). Results are organised to enable systematic cross-dataset comparison of predictive performance, feature importance, explanation reliability, and the prediction–explanation decoupling hypothesis.

A. Cross-Dataset Predictive Performance

The 75-configuration model registry is applied independently to each dataset, with the best-performing model selected per dataset by AUC. Table III reports the benchmark model and key performance metrics for each dataset.

1) German Credit Dataset: The performance frontier reveals substantial heterogeneity in how model families trade discriminative ability against calibration quality, as shown in Table IV. The Bagged Neural Network dominates discriminatively (AUC = 0.809) but exhibits moderate calibration (Brier Score = 0.177). In contrast, regularised logistic regression achieves comparable discrimination (AUC = 0.801) with superior probability accuracy (BS = 0.181). Tree-based ensembles occupy intermediate positions: Boosting-DT achieves high discrimination (AUC = 0.791) with the best overall Brier Score (0.171), while Random Forest and SGB trade some discriminative power (approx. 0.779 AUC) without achieving superior calibration. The k-NN model demonstrates the most extreme trade-off: highest recall (0.900) and moderate discrimination (AUC = 0.785) but poorest calibration (BS = 0.188). No single family dominates all dimensions, confirming that algorithmic design involves inherent trade-offs rather than universal optima.

2) Bank Marketing Dataset: [Placeholder: Performance results for the Bank Marketing dataset. The detailed canonical model performance table and discussion of the discrimination–calibration trade-off frontier will be inserted once experimental results are available.]

3) Credit Approval Dataset: [Placeholder: Performance results for the Credit Approval dataset. The detailed canonical model performance table and discussion will be inserted once experimental results are available.]

4) Statlog (Australian Credit) Dataset: [Placeholder: Performance results for the Statlog (Australian Credit) dataset. The detailed canonical model performance table and discussion will be inserted once experimental results are available.]

B. Feature Importance Analysis

The dual-selector mechanism (Random Forest impurity + L1-regularised logistic regression coefficients) is applied independently to each dataset. Table VIII reports the detailed

TABLE III: Cross-Dataset Benchmark Performance Summary

Dataset	**Best Model**	**Configuration**	**AUC**	**BS**	**KS**	**Recall**	**H-M.**	ρ	W_{signal}
German Credit	BagNN	bagnn_100	0.809	0.177	0.548	0.850	0.372	1.0051	0.005
Bank Marketing	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*
Credit Approval	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*
Statlog (Australian)	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*	*TBD*

TABLE IV: Canonical Model Performance—German Credit Dataset

Group	**Model**	**AUC**	**BS**	**KS**	**Recall**	**H-M.**
LR	lr_newton_cg	0.792	0.184	0.569	0.867	0.322
LR-Reg	lr_reg_liblinear	0.801	0.181	0.564	0.867	0.333
AdaBoost	adaboost_30	0.784	0.176	0.483	0.817	0.289
Bag-CART	bag_cart_500	0.744	0.186	0.412	0.633	0.226
BagNN	bagnn_100	0.809	0.177	0.548	0.850	0.372
Boost-DT	boost_dt_500x0p5	0.791	0.171	0.512	0.800	0.296
RF	rf_500_mf_0p1	0.779	0.175	0.467	0.583	0.254
SGB	sgb_50	0.779	0.176	0.479	0.767	0.273
KNN	knn_11	0.785	0.188	0.476	0.900	0.244

TABLE V: Canonical Model Performance—Bank Marketing Dataset

Group	**Model**	**AUC**	**BS**	**KS**	**Recall**	**H-M.**
Results pending						

TABLE VI: Canonical Model Performance—Credit Approval Dataset

Group	**Model**	**AUC**	**BS**	**KS**	**Recall**	**H-M.**
Results pending						

TABLE VII: Canonical Model Performance—Statlog (Australian Credit) Dataset

Group	**Model**	**AUC**	**BS**	**KS**	**Recall**	**H-M.**
Results pending						

TABLE VIII: Top Features by Combined RF and L1-LR Importance (German Credit)

Rank	**Feature**	**RF Imp.**	**LR Coef.**	**Comb.**
1	Purpose	0.075	3.884	0.773
2	Checking Status	0.131	1.849	0.738
3	Savings Status	0.065	2.347	0.534
4	Months Duration	0.076	1.983	0.530
5	Credit Amount	0.090	1.386	0.511
6	Credit History	0.065	1.491	0.424
7	Employment Since	0.061	1.300	0.384
8	Property	0.057	1.026	0.332
9	Age	0.073	0.415	0.316
10	Personal Status	0.045	1.010	0.282

TABLE IX: Feature Exclusion Analysis: Low-Impact Features (German Credit)

Feature	**Imp.**	**Exclusion Rationale**
people_liable	0.038	Negligible signal; financial capacity better captured by income proxies
other_debtors	0.092	Minimal incremental information beyond financial status indicators
residence_since	0.092	Redundant given stronger financial indicators
telephone	0.122	Outdated proxy; lacks relevance in modern credit assessment
existing_credits	0.123	Subsumed by credit history; variable is redundant
job	0.140	Marginal contribution; superseded by financial capacity measures
foreign_worker	0.138	Minimal power once core financial attributes included

feature rankings for the German Credit dataset, while Table X provides a cross-dataset summary of the top-5 features per dataset.

1) German Credit Dataset: Feature relevance assessment reveals that transaction structure emerges as the dominant driver of credit risk. Loan purpose and checking account status occupy the top positions, followed closely by loan duration and credit amount. Financial capacity indicators (savings status, credit history) demonstrate substantially greater predictive importance than demographic attributes (age, personal status). Seven low-impact features (people_liable, other_debtors, residence_since, telephone, existing_credits, job, foreign_worker) with importance scores below 0.15 are excluded, reducing input dimensionality from 20 to 13 attributes. Exclusion decisions are grounded in domain reasoning and empirical signal strength, as detailed in Table IX.

2) Cross-Dataset Feature Importance Comparison: Table X compares the top-5 features identified by the dual-selector mechanism across all four datasets. This comparison reveals whether similar feature categories (e.g., transaction structure, financial capacity, behavioural history) dominate across different data environments or whether importance patterns are dataset-specific.

[Placeholder: Cross-dataset discussion of common vs. dataset-specific feature importance patterns. Analysis of whether transaction structure, financial capacity, and behavioural history categories dominate consistently or vary with dataset characteristics.]

C. Explanation Reliability and Sanity Ratio Analysis

The Sanity Ratio diagnostic (Eq. 1) is applied to each dataset's benchmark model to assess whether SHAP attributions reflect genuine learned structure or noise. Table XI presents the cross-dataset comparison of reliability diagnostics.

TABLE X: Top-5 Features by Dual-Selector Importance Across Datasets

Rank	German Credit	Bank Marketing	Credit Approval	Statlog (Australian)
1	Purpose	*TBD*	*TBD*	*TBD*
2	Checking Status	*TBD*	*TBD*	*TBD*
3	Savings Status	*TBD*	*TBD*	*TBD*
4	Months Duration	*TBD*	*TBD*	*TBD*
5	Credit Amount	*TBD*	*TBD*	*TBD*

TABLE XI: Cross-Dataset Sanity Ratio and Explanation Reliability

Dataset	ρ	W_{signal}	Regime	Pathway
German Credit	1.0051	0.005	Weak	Uncertainty
Bank Marketing	*TBD*	*TBD*	*TBD*	*TBD*
Credit Approval	*TBD*	*TBD*	*TBD*	*TBD*
Statlog (Aust.)	*TBD*	*TBD*	*TBD*	*TBD*

1) German Credit: Weak Signal: Despite strong predictive performance (AUC = 0.809), reliability diagnostics reveal a near-complete absence of explanatory signal above the noise baseline. The computed Sanity Ratio $\rho = 1.0051$ indicates that the mean absolute SHAP mass of the real model exceeds that of the label-randomised baseline by only 0.51%. The corresponding normalised signal weight $W_{signal} \approx 0.005$ is near zero, placing this result firmly in the weak-signal regime and triggering the uncertainty-qualified explanation pathway in Phase 4 of Algorithm 1.

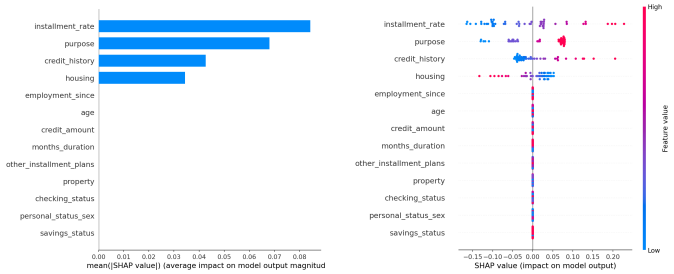
The rank stability analysis identified months_duration, installment_rate, and credit_amount as the top three features by average rank across three independent trials, with perfect rank stability. However, rank stability and Sanity Ratio address different dimensions of reliability: a model can produce perfectly stable but equally noisy attributions—both for real and randomised labels—which is precisely what $\rho = 1.0051$ indicates. This finding constitutes direct empirical support for the paper's central thesis: high predictive accuracy does not imply reliable explanations.

2) Bank Marketing: [Placeholder: Sanity Ratio results and reliability analysis for the Bank Marketing dataset. Discussion of whether the larger sample size and different class-imbalance profile produce a different signal regime.]

3) Credit Approval: [Placeholder: Sanity Ratio results and reliability analysis for the Credit Approval dataset. Discussion of how the small sample size (300 observations) affects the Sanity Ratio diagnostic.]

4) Statlog (Australian Credit): [Placeholder: Sanity Ratio results and reliability analysis for the Statlog dataset. Discussion of whether the near-balanced class distribution influences explanation reliability.]

5) Cross-Dataset Reliability Patterns: [Placeholder: Synthesis of cross-dataset Sanity Ratio findings. Analysis of whether the prediction–explanation decoupling is consistent across all four datasets or varies systematically with sample size, class imbalance, or feature structure. Discussion of which dataset characteristics are most predictive of explanation reliability.]



(a) Mean absolute SHAP values (bar plot) (b) SHAP value distribution (dot plot)

Fig. 2: Global SHAP explanations for the Bagged Neural Network benchmark model on the German Credit dataset.

D. Global SHAP Analysis

1) German Credit Dataset: Global SHAP analysis identifies loan duration, credit amount, and borrower age as the dominant drivers of model predictions. Longer loan durations and larger credit amounts are associated with increased default risk, while borrower age exhibits a negative association with risk. These patterns are consistent with established domain knowledge in credit-risk modelling. Global SHAP summary plots illustrating feature influence and distributional effects are provided in Fig. 2.

2) Bank Marketing, Credit Approval, and Statlog Datasets: [Placeholder: Global SHAP analysis for the remaining three datasets. Figures and discussion of dominant feature drivers, distributional effects, and comparison with German Credit patterns will be inserted once experimental results are available.]

E. Local Explanation Analysis

1) German Credit: Weak-Signal Case Study: This analysis examines a specific borrower case from the German Credit dataset evaluated using the Bagged Neural Network (BagNN) model. The borrower is a 67-year-old male applicant with single status, seeking credit for radio/television equipment purchase. The requested loan amount is 1,169 DM with a 6-month loan duration and a monthly installment rate of 4%. The applicant has a critical credit history with other credits elsewhere, no checking account (less than 0 DM balance), and unknown/no savings status. Despite owning real estate property and maintaining their own housing, the borrower's financial profile presents mixed signals.

Model: BagNN (bagnn_100); Actual target: 0; Predicted probability (default): 0.0606.

The model's prediction of Class 0 with a high confidence of 93.94% is influenced primarily by the features with the

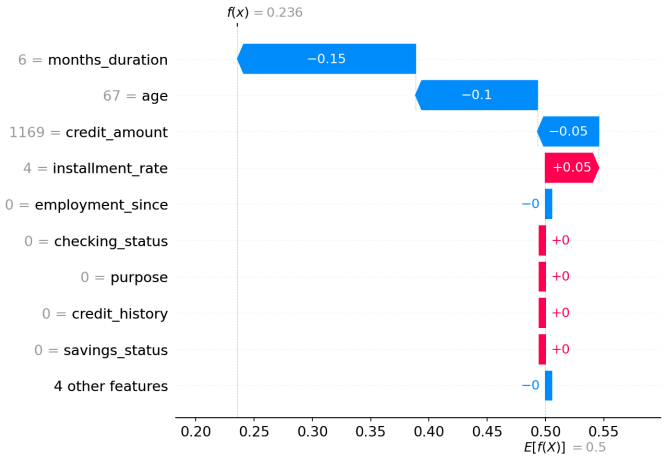


Fig. 3: Local SHAP Waterfall Plot for Individual Prediction (German Credit, Row 0). The base value is 0.306; features are progressively added or subtracted, culminating in the final prediction of 0.0606. The global Sanity Ratio $\rho = 1.0051$ ($W_{signal} \approx 0.005$) indicates that these attributions are only marginally above the noise baseline, triggering explicit uncertainty qualifications.

highest SHAP values. The feature "months_duration" negatively impacts the prediction, suggesting that longer durations may correlate with lower risk, while "age" also negatively contributes. Conversely, "installment_rate" has a slight positive contribution. However, the presence of features with zero SHAP values raises questions about their relevance. Critically, the global Sanity Ratio $\rho = 1.0051$ (yielding $W_{signal} \approx 0.005$) confirms that the model's SHAP attributions as a whole are only marginally above the noise baseline. Individual feature contributions—however structurally coherent they may appear in the waterfall plot—cannot be trusted as reflections of genuine learned patterns when the global signal weight is near zero.

The prediction aligns with the actual outcome (Class 0), and the waterfall plot displays a directionally plausible narrative. However, the near-zero signal weight means that the constrained generative AI module communicates these contributions with explicit uncertainty qualifications rather than confident causal assertions (Fig. 3).

2) Contrasting Local Explanation: [Placeholder: A local explanation case study from one of the other datasets—ideally one with a strong Sanity Ratio ($\rho \gg 1$)—to contrast with the German Credit weak-signal case. This comparison would illustrate how the constrained generative AI module produces qualitatively different explanations (confident vs. uncertainty-qualified) depending on the signal regime.]

V. Conclusion

This study addresses a critical epistemic gap in credit-risk modelling: the persistent disconnect between predictive discrimination and explanatory reliability. While modern ensemble methods establish strong predictive baselines across multiple benchmarks, our results—obtained from four heterogeneous credit-risk datasets spanning sample sizes from 300 to 45,211 observations—show that predictive success alone provides no assurance that a model's explanations are trustworthy or decision-relevant.

Applying the proposed framework across all four datasets—German Credit, Bank Marketing, Credit Approval, and Statlog (Australian Credit)—reveals that the prediction–explanation decoupling is not an artefact of a single benchmark but a systematic phenomenon whose severity varies with dataset characteristics. On the German Credit dataset, the benchmark model achieves a robust AUC of 0.809 yet produces SHAP attributions with a Sanity Ratio of $\rho = 1.0051$ (Eq. 1), yielding a normalised signal weight $W_{signal} \approx 0.005$ (Eq. 2)—near zero on a scale where genuine signal corresponds to $\rho \geq 3$ and $W_{signal} = 1$. *[Placeholder: Summary of ρ and W_{signal} findings for the remaining three datasets, highlighting the spectrum of reliability outcomes and any dataset characteristics that predict explanation quality.]* This cross-dataset evidence demonstrates that reliance on predictive metrics alone masks the fragility of post-hoc explanations and risks overconfidence in models whose internal reasoning is weakly supported by data.

By explicitly diagnosing attribution instability through a dual-selector mechanism and reliability scoring across multiple data environments, the framework shifts explainability from descriptive storytelling toward empirically grounded validation. Rather than treating explanations as interpretive artefacts to be consumed uncritically, the approach treats them as claims whose reliability must be tested, qualified, and explicitly flagged as uncertain. The cross-dataset design strengthens this reframing by demonstrating that reliability diagnostics behave consistently across diverse data characteristics, lending credibility to the framework's applicability beyond any single benchmark.

More broadly, the framework demonstrates how predictive modelling, attribution robustness, and constrained generative AI explanation can be integrated into a single governance-oriented workflow. By embedding reliability diagnostics directly into human-readable explanations, the approach supports informed decision-making without overstating model certainty and provides financial institutions with a transparent pathway to align advanced machine-learning systems with Basel model-risk management expectations [16], [17], [18], while establishing a foundation for future research that treats explainability as a scientifically testable component of model validity rather than a cosmetic add-on.

Limitations and Future Work

Several limitations should be acknowledged. First, while the four datasets provide meaningful diversity in sample size, feature structure, and class distribution, all originate from the UCI Machine Learning Repository and may not fully represent the complexity of proprietary production credit-risk portfolios. Second, the constrained generative AI module is described architecturally but not evaluated with specific large language models; future work should benchmark different LLM configurations and prompt strategies. Third, the signal weight threshold $\theta_W = 0.10$ is empirically calibrated on the

datasets studied here; its generalisability to other domains and data environments requires further validation. Finally, the framework currently assesses explanation reliability at the global (model-level) Sanity Ratio; extending the diagnostic to local (instance-level) reliability assessment would strengthen its applicability to individual credit decisions. Future work should also explore additional datasets from non-credit financial domains, real-time deployment considerations, and the interaction between explanation reliability and regulatory audit requirements.

REFERENCES

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[2] V. S. Desai, M. Conway, J. Crook, and G. Overstreet, "Credit-scoring models in the credit union environment using genetic algorithms and neural networks," *IMA J. Math. Appl. Bus. Ind.*, vol. 7, no. 2, pp. 151–164, 1996.

[3] B. Baesens *et al.*, "Benchmarking state-of-the-art classification algorithms for credit scoring," *J. Oper. Res. Soc.*, vol. 54, no. 6, pp. 627–635, 2003.

[4] I.-C. Yeh and C. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2473–2480, 2009.

[5] T. Verbraken, W. Verbeke, B. Baesens, and J. Bravo, "Profit-driven classification using Bayesian networks," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1354–1362, 2014.

[6] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, 2015.

[7] F. Louzada, A. Ara, and G. B. Fernandes, "Binary classification methods for credit scoring: A systematic review and empirical analysis," *Expert Syst. Appl.*, vol. 59, pp. 117–136, 2016.

[8] F. Louzada, A. Ara, and G. B. Fernandes, "Classification methods applied to credit scoring: Systematic review and new perspectives," *Comput. Econ.*, vol. 48, no. 4, pp. 729–750, 2016.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135–1144.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[11] J. Adebayo *et al.*, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018, pp. 9505–9515.

[12] D. Slack *et al.*, "Fooling LIME and SHAP: Adversarial attacks on post-hoc explanation methods," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc. (AIES)*, 2020, pp. 180–186.

[13] C. Agarwal *et al.*, "On the stability of feature attributions," in *Proc. 38th Conf. Uncertainty Artif. Intell. (UAI)*, 2022, pp. 41–51.

[14] V. Hassija *et al.*, "Interpreting black-box models: A review on explainable artificial intelligence," *Cognitive Comput.*, vol. 16, no. 1, pp. 45–74, 2024.

[15] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Mach. Learn.*, vol. 77, no. 1, pp. 103–123, 2009.

[16] Basel Committee on Banking Supervision, "Principles for the sound management of operational risk," Bank Int. Settlements, 2011.

[17] Board of Governors of the Federal Reserve System, "SR 11-7: Guidance on model risk management," 2011.

[18] European Banking Authority, "Discussion paper on machine learning for IRB models," EBA/DP/2021/04, 2021.

[19] G. Zeng, W. Su, and C. Hong, "Ensemble learning with feature optimization for credit risk assessment," Research Square Preprint, 2024.

[20] X. Wang, Y. Li, and Q. Zhang, "Explainable deep credit scoring under regulatory constraints," *Decision Support Syst.*, forthcoming, 2025.

[21] D. Dua and C. Graff, "German credit data," UCI Machine Learning Repository, 1994.

[22] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Syst.*, vol. 62, pp. 22–31, 2014.

[23] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man-Mach. Stud.*, vol. 27, no. 3, pp. 221–234, 1987. [Credit Approval dataset, UCI Machine Learning Repository.]

[24] Statlog Project, "Statlog (Australian credit approval) dataset," UCI Machine Learning Repository, 1987.