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 regularization; (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 on the German Credit dataset shows that high-performing models can produce feature attributions indistinguishable from noise, establishing empirical grounding for the falsifiability approach. 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 [**?**], [**?**]. 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 [**?**]. 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 [**?**], [**?**]. 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 [**?**], [**?**]. 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 [**?**]. 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 reasoning while explicitly quantifying whether explanatory signals exceed noise baselines under empirical scrutiny [**?**], [**?**].

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 Optimization Framework proposed by Zeng *et al.* [**?**]. Their dual-selector mechanism—combining nonlinear Random Forest importance with sparse L1-regularised logistic regression coefficients—establishes a 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.

Using the German Credit dataset as a controlled benchmark,

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

the framework is evaluated across a broad family of calibrated classification models to assess both predictive performance and the empirical reliability of their explanations [**?**], [**?**]. The results reveal a structural paradox: explanation quality varies independently of predictive accuracy, and even high-performing models can produce feature attributions requiring careful reliability assessment. By providing explicit, quantifiable diagnostics for explanation reliability, 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 Optimization

Early credit-risk models relied on classical statistical techniques such as logistic regression and linear discriminant analysis, valued for transparent coefficient structures [**?**]. However, these approaches struggle to capture nonlinear interactions and heterogeneous borrower behaviour. Comparative benchmarks, notably by Baesens *et al.* [**?**], 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) [**?**], [**?**]. 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 [**?**], [**?**].

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 [**?**], [**?**]. Addressing this, Zeng *et al.* [**?**] proposed a Feature Selector-classifier Optimization 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 [**?**]. 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 [**?**], [**?**].

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 [**?**], SR 11-7 guidance [**?**], and EBA discussions on machine learning for internal ratings-based models [**?**] emphasise the need for transparent, auditable model explanations. While traditional models offered intrinsic interpretability [**?**], 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 [**?**] and SHAP [**?**] 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 [**?**]. 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.* [**?**] demonstrate that attribution scores often conflate signal and noise, while Slack *et al.* [**?**] 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 [**?**]. 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 [**?**], [**?**], attribution methods [**?**], [**?**], or generative explanation [**?**], [**?**]—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 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. **??** 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.

B. Data and Preprocessing

The experiments utilise the German Credit dataset from the UCI Machine Learning Repository [**?**], a widely used benchmark in credit-risk research comprising 1,000 observations (700 non-default and 300 default cases) and 20 attributes [**?**]. This dataset is chosen as a controlled benchmark precisely because of its moderate size and complexity: large enough to support ensemble learning yet small enough for thorough reliability assessment without computational barriers to exhaustive cross-validation and sanity checking. While this single-dataset design limits empirical generality, it enables thorough methodological validation of the falsifiability framework under controlled conditions.

To ensure robust model estimation and prevent confounding of predictive and explanatory reliability, the data 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-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.

Class imbalance is addressed using the Synthetic Minority Over-sampling Technique (SMOTE), applied separately within each training fold during stratified cross-validation. 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 [**?**], [**?**].

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 **??**. 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.

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

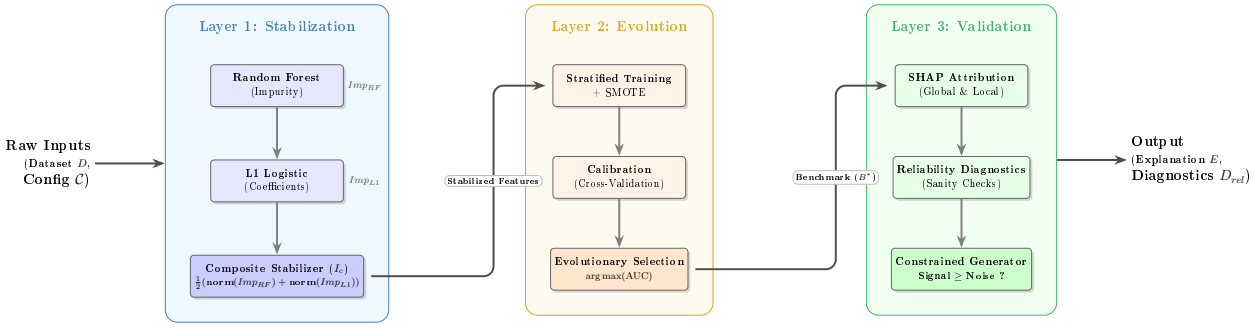


Fig. 1: Conceptual Architecture: Task Decomposition and Information Constraints.

TABLE I: 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**

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) Feature Attribution and Generative 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 module that translates quantitative attributions into human-readable narratives. Critically, the generative 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 **??**. 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).

Algorithm 1: Operational Instantiation: Unified Predictive and Explanatory Workflow

Input: Dataset D, Model Registry $\mathcal{M}$, Configuration Set $\mathcal{C}$
Output: Benchmark Model B^*, Explanations E, Reliability Diagnostics D_{rel}
Phase 1: Feature Screening (Dual-Selector)
Train Random Forest and L1-logistic regression; obtain importance scores Imp_{RF} and Imp_{L1}.
Compute composite score: $I_c = \frac{1}{2}(\text{norm}(Imp_{RF}) + \text{norm}(Imp_{L1}))$.
Rank features by I_c; apply domain-based exclusion rationale; select screened feature set.
Phase 2: Model Training and Selection
for *each model family* $F \in \mathcal{M}$ **do**
  Perform stratified split; apply SMOTE within each fold; train calibrated instantiations; evaluate AUC.
Define benchmark: $B^* = \arg\max_{B \in \bigcup_F \mathcal{M}_F}(\text{AUC}(B))$.
Phase 3: Explainability and Reliability Assessment
Compute SHAP values (global and local) for B^*.
Evaluate attribution stability across multiple trials; compute Sanity Ratio $\rho \in [0, 1]$, where $\rho \approx 1$ indicates stable attributions.
Set reliability threshold $\theta = 0.95$ (empirically calibrated)
Phase 4: Constrained Generative Explanation
Extract top-k SHAP features; construct structured prompt (feature names, values, SHAP contributions, prediction, label).
if $\rho \geq \theta$ *(high reliability)* **then**
  Invoke LLM to generate explanation grounded in validated SHAP evidence
else
  Invoke LLM to generate explanation with explicit uncertainty qualifications
return B^*, E, D_{rel}

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 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 [**?**]; 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, organised to distinguish between predictive performance, feature importance, and explanatory reliability.

A. Supervised Feature Importance Analysis

Feature relevance is assessed using the proposed dual-selector mechanism, which combines Random Forest impurity-based importance with coefficient magnitudes from L1-regularised logistic regression. The aggregated feature rankings are reported in Table **??**. Several key patterns emerge from the analysis.

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. These variables characterise the fundamental structure and terms of the transaction, suggesting that the nature and scope of the credit request are central to default risk assessment.

Financial capacity indicators demonstrate substantially greater predictive importance than demographic attributes. Savings status and credit history rank significantly higher than borrower age or personal characteristics, indicating that established financial behaviour and accumulated resources provide stronger signals of creditworthiness than static demographic properties. This finding aligns with established credit-risk theory, which emphasises the primacy of financial position over personal circumstances.

Behavioural history proves more informative than employment stability. Credit history substantially outweighs employment tenure in explaining risk, revealing that a borrower's track record of credit management carries greater explanatory

TABLE II: Top Features by Combined RF and L1-LR Importance

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 III: Feature Exclusion Analysis: Rationale for Low-Impact Features

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 (checking, savings)
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

power than tenure in current employment. This pattern underscores the importance of demonstrated financial discipline across the credit lifecycle.

Social indicators contribute minimally to discriminative power. Variables such as foreign worker status, telephone ownership, and residential stability add negligible incremental signal beyond the stronger indicators already identified. These low-impact features are subsequently excluded from the modelling phase, reducing input dimensionality without meaningful loss of predictive information.

B. Feature Selection Rationale

Features are ranked by dual-selector importance and systematically evaluated for inclusion. Seven low-impact features with importance scores below 0.15 are excluded from modelling. Exclusion decisions are grounded in domain reasoning and empirical signal strength, as detailed in Table **??**.

Excluding these features reduces input dimensionality from 20 to 13 attributes, retaining only those variables that collectively drive credit-risk discrimination while minimising noise and model instability.

C. Canonical Model Instantiations and Performance

Canonical instantiations are selected per family based on theoretical and structural principles. This approach ensures that comparative assessment reflects genuine algorithmic differences rather than configuration-specific tuning effects.

Linear Models are instantiated through two complementary variants. Unregularised logistic regression with the newton-cg solver provides the classical statistical baseline, eschewing explicit complexity penalties to preserve the foundational logistic model. Regularised logistic regression with L1-penalty via liblinear represents the modern variant that intrinsically enforces sparsity, automatically excluding irrelevant features through the penalty mechanism.

Boosting is represented by AdaBoost with decision stumps (ensemble size 30). This instantiation embodies the original boosting principle grounded in iterative error correction. The shallow base learner (one-level decision tree) ensures theoretical clarity and computational transparency, avoiding the confounding effects of deeper trees or more complex base learners.

Bagging (CART) employs bagged decision trees with ensemble size 500, instantiating classical bootstrap aggregation without feature subsampling. This configuration establishes the baseline ensemble principle, serving as a reference point for understanding the effects of feature subsampling introduced in Random Forest variants.

Bagging (Neural Network) extends the bagging principle to flexible nonlinear function approximators. Bagged neural networks with ensemble size 100 represent the intersection of modern deep-learning architectures with classical ensemble methodology, demonstrating how bagging performs when applied to highly flexible learners rather than tree-based models.

Random Forest instantiation employs $\sqrt{p}$ feature subsampling (ensemble size 500), representing a specialised variant that introduces controlled feature randomness. This configuration represents a key theoretical contribution to ensemble learning, trading reduced per-tree correlation for improvements in generalisation through feature decorrelation.

Stochastic Gradient Boosting (ensemble size 50, learning rate 0.1) instantiates the gradient-boosting principle with conservative operational parameters. The modest ensemble scale and conservative learning rate reflect a balanced approach that prioritises approximation quality and generalisation stability over aggressive boosting.

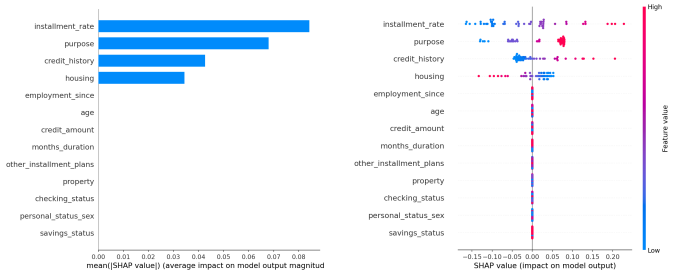
Instance-Based methods are represented by k-NN with $k = 11$ and uniform distance weighting. This neighbourhood size balances local responsiveness (small k overfits to individual observations) against global smoothing (large k ignores local structure), establishing the non-parametric neighbourhood principle at an intermediate operational point.

The performance frontier reveals substantial heterogeneity in how model families trade discriminative ability against calibration quality, as shown in Table **??**. 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). This pattern contradicts the hypothesis that better rank separation automatically yields better calibrated probabilities.

Tree-based ensembles occupy intermediate positions:

TABLE IV: Canonical Model Instantiations—Performance Metrics on 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



(a) Mean absolute SHAP values (bar plot)

(b) SHAP value distribution (summary dot plot)

Fig. 2: Global SHAP explanations for the Bagged Neural Network benchmark model on the German Credit dataset. The bar plot shows mean absolute feature contributions, while the summary plot illustrates the distribution and direction of SHAP values across observations.

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).

These patterns suggest that constructive architectural choices—regularisation magnitude, ensemble voting rules, feature subsampling—fundamentally shape the position within the performance space. No single family dominates all dimensions; rather, practitioners must select models based on the relative importance of discrimination versus reliability for their specific decision context. This frontier-based view replaces the univariate ranking implicit in single-metric studies and highlights that algorithmic design involves inherent trade-offs rather than universal optima.

D. Global Explainability Analysis

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. **??**.

E. Feature Stability and Sanity Validation

To assess the reliability and consistency of SHAP-based explanations, we conducted a feature stability analysis using three trials with a background size of 50. The Sanity Ratio of 0.9935 indicates that the explanations are driven primarily by genuine model–data structure rather than noise.

The analysis identified the top three features by average rank: months_duration, installment_rate, and credit_amount. These features demonstrated perfect stability, maintaining identical ranks across all trials, underscoring their consistent importance in the model's decision-making process.

Among the remaining ten features, stability varied considerably. Most features exhibited stable rankings, while some showed moderate variation and others exhibited unstable rankings. The sanity ratio of 0.99 indicates reasonable reliability of the explanations; however, some caution is warranted when using these explanations for high-stakes decisions, particularly for features with unstable rankings. This finding emphasises the importance of validating explanation stability beyond raw predictive performance metrics.

F. Explanation Reliability

Despite strong predictive performance, reliability diagnostics reveal substantial weaknesses in explanatory stability. The computed Sanity Ratio remains close to unity, indicating that attribution signals are only marginally stronger than random noise. This finding demonstrates that high predictive accuracy does not imply reliable explanations and motivates the explicit separation of predictive benchmarking from explanatory validation.

G. Local Explanation Analysis

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: the lack of established checking and savings accounts suggests limited financial footprint, while the property ownership indicates some asset base.

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 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, indicating that older individuals might be perceived as lower risk. Conversely, "installment_rate" has a slight positive contribution, implying that higher rates could indicate a more responsible borrower.

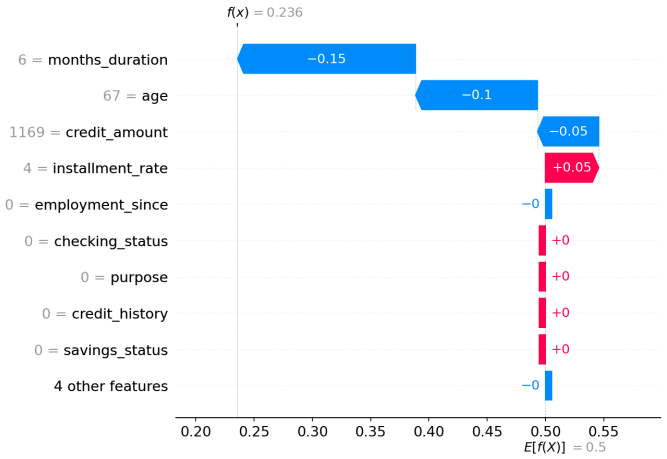


Fig. 3: Local SHAP Waterfall Plot for Individual Prediction (Row 0). The base value (model's average output) is 0.306; features are progressively added or subtracted based on their SHAP values, culminating in the final prediction of 0.0606 (non-default). The dominance of months_duration and age in moving the prediction toward non-default is evident. The weak signal-to-noise ratio indicates these contributions should be communicated with explicit uncertainty qualifications.

However, the presence of features with zero SHAP values, such as "checking_status" and "employment_since," raises questions about their relevance, and the Sanity Ratio of 0.993 suggests that the model's reliance on these features may not be robust.

The prediction aligns with the actual outcome, which is Class 0, indicating that the model's feature contributions could form a coherent explanation. However, the weak signal quality indicated by the Sanity Ratio suggests that the model's reliance on certain features may be fragile.

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 such as Bagged Neural Networks (BagNN) and Boosting establish strong predictive baselines in standard benchmarks, our results show that predictive success alone provides no assurance that a model's explanations are trustworthy or decision-relevant.

Applying the proposed unified predictive–explanatory framework reveals a structural paradox at the core of contemporary explainable AI practice. Despite achieving robust AUC scores ($>$0.80), many models produce feature attributions with Sanity Ratios close to 1.015, indicating explanatory signals barely distinguishable from random noise. This 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. In practice, explanation quality varies independently of predictive accuracy.

By explicitly diagnosing attribution instability through a dual-selector mechanism and reliability scoring, 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. This reframing is essential for regulated credit-risk environments, where transparency, challengeability, and auditability are as important as predictive performance.

More broadly, the framework demonstrates how predictive modelling, attribution robustness, and constrained generative 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, while establishing a foundation for future research that treats explainability as a scientifically testable component of model validity rather than a cosmetic add-on.

Acknowledgment

We acknowledge the computational support and data resources that enabled this research. We thank colleagues for thoughtful feedback on earlier drafts.

References

[1] W. H. Beaver, "Financial ratios as predictors of failure," *J. Accounting Res.*, vol. 4, no. 1, pp. 71–111, 1966.
[2] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *J. Finance*, vol. 23, no. 4, pp. 589–609, 1968.
[3] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *J. Accounting Res.*, vol. 18, no. 1, pp. 109–131, 1980.
[4] J. C. Wiginton, "A note on the comparison of logit and discriminant models of consumer credit behavior," *J. Financial Quant. Anal.*, vol. 15, no. 3, pp. 757–770, 1980.
[5] L. C. Thomas, D. B. Edelman, and J. N. Crook, *Credit Scoring and Its Applications*. Philadelphia, PA, USA: SIAM, 2002.
[6] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.
[8] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals Syst.*, vol. 2, no. 4, pp. 303–314, 1989.
[9] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.
[10] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
[13] 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.
[14] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
[16] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," *arXiv preprint arXiv:2012.06678*, 2020.
[17] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 18932–18943.

[18] V. Borisov *et al.*, "Deep neural networks and tabular data: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 4–26, 2022.

[19] 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.

[20] 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.

[21] C.-F. Tsai and J.-W. Wu, "Using neural network ensembles for bankruptcy prediction and credit scoring," *Expert Syst. Appl.*, vol. 34, no. 4, pp. 2639–2649, 2008.

[22] 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.

[23] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit risk models via machine-learning algorithms," *J. Banking Finance*, vol. 34, no. 11, pp. 2767–2787, 2010.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] S. Bach *et al.*, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, Art. no. e0130140, 2015.

[29] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Statist.*, vol. 24, no. 1, pp. 44–65, 2015.

[30] 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.

[31] 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.

[32] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.

[33] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *J. Roy. Statist. Soc. Ser. B*, vol. 82, no. 4, pp. 1059–1086, 2020.

[34] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

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

[36] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in *Proc. 2018 ICML Workshop Human Interpretability Mach. Learn.*, 2018, pp. 66–71.

[37] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.

[38] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[39] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[40] R. Caruana *et al.*, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. 21st ACM SIGKDD*, 2015, pp. 1721–1730.

[41] 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.

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

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

[44] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[45] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[46] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[47] 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.

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

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

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

[51] G. Yildirim and M. O. Kulekci, "Graph neural networks for credit risk analysis," *Expert Syst. Appl.*, vol. 186, Art. no. 115822, 2021.

[52] X. Li and Y. Wu, "Advanced post-hoc interpretability in financial modelling," *J. Financial Data Sci.*, vol. 6, no. 1, pp. 10–25, 2024.

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

[54] J. Quan and X. Sun, "Credit risk assessment using the factorization machine model with feature interactions," *Humanities Social Sci. Commun.*, vol. 11, no. 234, 2024.

[55] C. Wang, K. Zhang, and H. Wang, "Interpretable credit risk modelling: Foundations, challenges, and future directions," *Decision Support Syst.*, vol. 181, Art. no. 113902, 2025.

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

[57] L. Wang, Z. Yu, J. Ma, X. Chen, and C. Wu, "A two-stage interpretable model to explain classifier in credit risk prediction," *J. Forecasting*, 2025.

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