International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 3, May-June 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Towards Interpretable Credit Risk Assessment: A Comparative Study of XAI Techniques with Regulatory Compliance

Charvak Dharmaraj Chavare[1], Chaitanya Ramesh Darekar[2]
[1](Department Computer Science, Viday Pratishthans's Art, Science and Commerce, and Baramati
Email: charvakchaware@gmail.com)
[2](Department Computer Science, Viday Pratishthans's Art, Science and Commerce, and Baramati
Email: chaitanyadarekar2002@gmail.com)

----------**********----------

Abstract:

Credit risk assessment increasingly relies on high-performance black-box ML models such as XGBoost and deep neural networks. While these models deliver superior predictive accuracy, their opacity conflicts with regulatory frameworks comprising GDPR's right-to-explanation and fair lending laws. This paper applies and compares three post-hoc XAI techniques; SHAP, LIME, and counterfactual clarifications; on various and several black-box models trained on a real-world credit dataset (for example, German Credit, HELOC, LendingClub). We evaluate each strategy on faithfulness, stability, computational cost, and human interpretability. Results show that SHAP offers the most globally uniform and unchanging clarifications, while counterfactuals are most actionable for individual loan applicants. We further discuss implications for regulatory adherence and propose a decision-support framework for deploying explainable credit scoring in practice.

Keywords — **Explainable AI (XAI), Credit Risk Assessment, Black-Box Models, SHAP & LIME, Counterfactual Explanations, Regulatory Compliance**

----------**********----------

1: Introduction

1.1 Background

The financial sector has undergone a deep and intense transformation over the past decade, driven by the quick adoption of machine learning (ML) and artificial intelligence (AI) in core decision-making operations. Among these, credit risk assessment; the process of assessing the likelihood that a borrower will default on a loan; has seen a notable shift from customary statistical models such as logistic regression and scorecards toward high-performance black-box models comprising gradient boosting machines, random forests, and deep neural networks. These models consistently outperform their classical counterparts on predictive accuracy, handling non-linear connections and high-dimensional feature spaces with exceptional and striking productivity.

However, this performance increase comes at an essential cost: interpretability. Black-box models, by their very nature, do not readily uncover the reasoning behind their predictions. In a domain as consequential as credit lending, where a denied loan can materially influence a person's financial life, the incapacity to clarify a model's decision is not merely an academic concern; it is a legal, moral, and regulatory one.

1.2 Problem Statement

Despite the superior predictive abilities of modern ML models, their extensive and prevalent adoption in credit risk assessment is hindered by a core and foundational tension between accuracy and transparency. Regulators, auditors, and impacted individuals increasingly require that automated financial decisions be explainable, auditable, and fair. Financial organizations that deploy opaque ML models risk non-compliance

with regulatory mandates, erosion of customer trust, and the perpetuation of discriminatory lending practices embedded within training data.
The core problem this inquiry addresses is: How can black-box ML models used in credit risk assessment be made comprehensible without considerably sacrificing predictive performance, and which Explainable AI (XAI) technique best serves the necessities of regulated financial environments?

1.2 Problem Statement

Despite the superior predictive abilities of modern ML models, their extensive and prevalent adoption in credit risk assessment is hindered by a core and foundational tension between accuracy and transparency. Regulators, auditors, and influenced individuals increasingly require that automated financial decisions be explainable, auditable, and fair. Financial organizations that deploy opaque ML models risk non-compliance with regulatory mandates, erosion of customer trust, and the perpetuation of discriminatory lending practices embedded within training data.
The core problem this study addresses is: How can black-box ML models used in credit risk assessment be made comprehensible without considerably sacrificing predictive performance, and which Explainable AI (XAI) technique best serves the necessities of regulated financial environments?

1.3 Motivation

Several converging forces motivate this inquiry:
Regulatory pressure is mounting globally. The European Union‘s General Data Protection Regulation (GDPR), specifically Article 22, grants individuals the right not to be subject to entirely automated decisions that considerably influence them, and the right to obtain a significant and purposeful explanation of such decisions. The EU AI Act (2024) further classifies credit scoring as a high-risk AI application, mandating transparency and human oversight. In the United States, the Equal Credit Opportunity Act (ECOA) and Fair Housing Act need lenders to provide precise and particular reasons for harmful and detrimental credit decisions. Basel III further reinforces the need for model risk management.
Ethical concerns around algorithmic bias are equally pressing. ML models trained on historical lending data may encode systemic biases against shielded demographic groups, arising in discriminatory results that is challenging to detect without interpretability instruments. Practical need drives the need for explainability at both the worldwide level; understanding which features generally drive credit decisions; and the local level; explaining why a particular applicant was authorized or denied, and what they could adjust to obtain consent.

1.4 Objectives

This inquiry aims to achieve the following objectives:

1. To examine and categorise existing Explainable AI (XAI) techniques applicable to credit risk assessment, differentiating between worldwide and local, ante-hoc and post-hoc approaches.
2. To train and evaluate several black-box ML models; incorporating XGBoost, Random Forest, LightGBM, and a Multilayer Perceptron; on publicly available credit datasets.
3. To apply and compare three post-hoc XAI techniques; SHAP (SHapley Additive elucidations), LIME (Local Interpretable Model-agnostic Explanations), and DiCE (Diverse Counterfactual Explanations); across the trained models.
4. To evaluate each XAI strategy on quantitative criteria incorporating faithfulness, stability, and computational productivity, as well as qualitative criteria comprising regulatory alignment and human interpretability.
5. To propose a pragmatic framework for selecting and deploying XAI approaches in regulated credit lending environments.

1.5 Research Questions

This paper is led by the following inquiry queries:

- RQ1: Which black-box ML model achieves the best predictive performance on credit risk classification activities?
- RQ2: How do SHAP, LIME, and counterfactual clarifications differ in the quality, consistency, and actionability of clarifications they generate?
- RQ3: Which XAI approach is best aligned with the transparency necessities of financial regulators under GDPR, ECOA, and the EU AI Act?
- RQ4: Is there a trade-off between model accuracy and explainability quality, and how can it be managed in practice?

1.6 Scope and Limitations

This study concentrates exclusively on binary credit default classification utilizing organized, tabular financial data. The datasets used; the German Credit Dataset (UCI Repository), the HELOC dataset (FICO), and a subset of the LendingClub dataset; are publicly available benchmarks broadly used in academic literature, guaranteeing reproducibility. The inquiry does not extend to unstructured data such as text-based credit reports or image-based identity validation.

The XAI strategies assessed are limited to post-hoc, model-agnostic techniques. Ante-hoc understandable models such as Explainable Boosting Machines (EBM) and Neural Additive Models (NAM) are discussed in the literature review but not experimentally assessed, representing a direction for future work. Additionally, real-world deployment restrictions such as model serving latency and integration with legacy banking system are beyond the extent of this study.

1.7 Organisation of the Report

The remainder of this paper is organised as follows. Chapter 2 presents an exhaustive and inclusive literature review covering the evolution of credit scoring, XAI taxonomies, and the regulatory landscape. Chapter 3 particulars the inquiry methodology comprising datasets, model training, XAI application, and evaluation criteria. Chapter 4 presents and analyses experimental outcomes. Chapter 5 discusses results in the setting of regulatory adherence and pragmatic deployment. Chapter 6 concludes the paper and outlines directions for future study.

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.

Recommended font sizes are shown in Table 1.

2: Literature Review

2.1 Introduction to the Chapter

This chapter presents a complete and thorough review of existing literature across three interconnected domains: the evolution of credit risk assessment models, the taxonomy and techniques of Explainable AI (XAI), and the regulatory frameworks governing automated financial decision-making. The review identifies fundamental contributions, systematic trends, and inquiry gaps that this study aims to address.

2.2 Evolution of Credit Risk Assessment

2.2.1 Traditional Statistical Models

Credit risk assessment has a long history rooted in statistical and actuarial techniques. The earliest formalised approach was the credit scorecard, developed in the 1950s by Fair, Isaac and Company (FICO), which assigned numerical scores to applicants established on weighted linear combinations of financial attributes such as payment history, credit utilisation, and length of credit history (Hand & Henley, 1997). These scorecards, built primarily applying logistic regression and linear discriminant analysis, provided high interpretability and regulatory approval, and persist in application today.

Altman (1968) introduced the Z-score model, a multivariate discriminant analysis technique for forecasting corporate bankruptcy employing five financial ratios. Though designed for corporate credit risk, it founded the fundamental rule of

integrating numerous and manifold financial indicators into a single predictive score. Subsequent decades observed enhancements through probit models, survival analysis, and danger models, each contributing greater statistical rigour but functioning within the identical comprehensible, linear paradigm (Thomas et al., 2002).

The primary limitation of these customary models lies in their assumption of linearity and incapacity to capture intricate, non-linear interactions between features; a notable constraint given the multidimensional and frequently non-linear nature of borrower behaviour.

2.2.2 Machine Learning in C
redit Scoring

The arrival of machine learning brought a paradigm shift in credit risk modelling. Decision trees, introduced by Breiman et al. (1984) through the CART algorithm, provided non-linear decision boundaries and natural interpretability, making them an initial favourite in the financial sector. However, individual decision trees are prone to overfitting and instability. Ensemble approaches addressed these limitations considerably. Random Forest (Breiman, 2001), an ensemble of decorrelated decision trees trained via bagging, demonstrated superior generalisation and robustness. Gradient Boosting Machines (Friedman, 2001), and their optimised types XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), further advanced predictive performance and became the primary and principal paradigm in credit scoring competitions and sector deployments.

Support Vector Machines (SVM) were investigated thoroughly for credit classification due to their robust theoretical foundations and capability to handle high-dimensional feature spaces (Huang et al., 2007). Artificial Neural Networks (ANN) and Deep Learning models, comprising multilayer perceptrons and recurrent architectures, have moreover been applied to credit default prediction, demonstrating robust performance on big and substantial datasets (Wang et al., 2011; Kvamme et al., 2018). An exhaustive and inclusive comparative study by Lessmann et al. (2015) benchmarked 41 classification techniques across eight credit datasets, concluding that random forests and gradient boosting consistently outperformed conventional logistic regression, but at the cost of transparency. This accuracy-interpretability trade-off is the central tension inspiring the present study.

2.2.3 The Accuracy-Interpretability Trade-off

The shift toward high-performance black-box models in credit scoring has been well-documented in the literature. Rudin (2019) contends forcefully against the application of black-box models in high-stakes decisions, contending that comprehensible models can achieve comparable accuracy while remaining fully clear. However, empirical evidence proposes that for intricate, high-dimensional financial datasets, black-box models retain a significant and purposeful performance benefit (Baesens et al., 2003). This creates the central dilemma: sacrificing accuracy for transparency, or accepting opacity for performance. XAI emerges as the proposed resolution; implementing explanation strategies to black-box models post-hoc, maintaining both accuracy and transparency.

2.3 Explainable Artificial Intelligence (XAI)

2.3.1 Definition and Taxonomy

Explainable AI refers to a set of techniques and techniques that make the predictions and internal workings of ML models clear to human users. Arrieta et al. (2020) provide a complete and thorough survey and taxonomy of XAI, differentiating along two primary dimensions:

Scope of explanation:

- Global explainability- understanding the overall behaviour of a model across all predictions (for example, which features is most crucial and vital in broad and inclusive)
- Local explainability- understanding why a particular individual prediction was made (for example, why this applicant was denied)

Timing of explanation:

- Ante-hoc (inherent and essential) interpretability- models that is comprehensible by design, such as logistic regression, decision trees, and Explainable Boosting Machines (EBM)
- Post-hoc explainability- explanation approaches applied after training to black-box models, such as SHAP, LIME, and counterfactual clarifications

This study concentrates on post-hoc, model-agnostic strategies, as they can be applied uniformly across dissimilar model architectures without requiring architectural changes.

2.3.2 SHAP (SHapley Additive exPlanations)

SHAP, introduced by Lundberg & Lee (2017), is grounded in collaborative game theory, specifically Shapley values from Shapley (1953). SHAP computes the contribution of each feature to a prediction by considering all possible subsets of features and averaging their minimal and negligible contributions. The arising explanation fulfills three attractive attributes: local accuracy (elucidations sum to the model outcome), missingness (absent features have zero contribution), and consistency (features that contribute more invariably receive greater attribution).

Lundberg et al. (2020) extended SHAP with TreeSHAP, a productive and effective algorithm specifically optimised for tree-based models such as XGBoost and Random Forest, minimizing computation from exponential to polynomial time. SHAP has become the de facto standard for feature attribution in credit risk literature due to its theoretical rigour and consistency. Several studies have confirmed SHAP‘s effectiveness in credit scoring contexts. Bussmann et al. (2021) demonstrated that SHAP elucidations for loan default prediction models aligned strongly with domain specialist intuition, supporting regulatory audit operations. Gramegna & Giudici (2021) compared SHAP with other attribution approaches and found it to be the most stable and faithful across distinct model classes.

2.3.3 LIME (Local Interpretable Model-agnostic Explanations)

LIME, proposed by Ribeiro et al. (2016), generates local elucidations by fitting an uncomplicated and elementary, understandable surrogate model; normally a scant linear model; in the neighbourhood of the prediction instance being clarified. Perturbations are generated around the instance, predictions are obtained from the black-box model, and the surrogate is fitted on these perturbed samples weighted by their proximity to the initial instance.

LIME provides broad relevance across model sorts, data methods (tabular, content, image), and is computationally lightweight. However, it has been criticised for instability; repeated clarifications for the identical instance can differ considerably due to the stochastic nature of disturbance sampling (Alvarez-Melis & Jaakkola, 2018). Shankar et al. (2021) demonstrated that LIME clarifications in credit scoring contexts, while locally correct, exhibit greater variance than SHAP, making them less fitting for regulatory reporting where consistency is necessary and fundamental.

2.3.4 Counterfactual Explanations

Counterfactual clarifications, formalised by Wachter et al. (2017), address a distinct but complementary query: not “why was this decision made?” but “what would need to adjust for a distinct decision to be made?” A counterfactual explanation for a denied loan application might state: “If your annual earnings were ₹ 50,000 greater and your existing debt reduced by 20%, your application would have been authorized.” This form of explanation is especially valuable for actionability; offering rejected applicants with specific, realistic steps to improve their credit profile. It is furthermore closely aligned with the GDPR necessity for relevant and substantial explanation, as it offers recourse alternatively than merely attribution.

DiCE (Diverse Counterfactual Explanations), developed by Mothilal et al. (2020), generates various and several varied and distinct counterfactuals concurrently, offering applicants various alternative pathways. Studies have displayed that counterfactual clarifications are

preferred by end users over feature attribution approaches in user studies, though they are more computationally expensive to generate (Keane & Smyth, 2020).

2.3.5 Additional XAI Techniques

Partial Dependence Plots (PDP), introduced by Friedman (2001), visualise the minimal and negligible effect of one or two features on predicted results, averaged across all other features. They provide worldwide, aggregate-level insights but can be deceptive in the presence of feature correlations. Individual Conditional Expectation (ICE) plots (Goldstein et al., 2015) extend PDPs by showing the reliance curve for each individual instance alternatively than the sum, revealing varied effects that PDPs may obscure. Attention processes in neural networks have been proposed as a form of built-in interpretability, though Jain & Wallace (2019) demonstrated that attention weights do not dependably reflect feature significance, tempering their application as explanation instruments in credit contexts.

2.4 Regulatory Frameworks and Compliance Requirements

2.4.1 GDPR and the Right to Explanation

The General Data Protection Regulation (GDPR), enacted by the European Union in 2018, denotes the most notable regulatory development for AI-driven credit decisions in latest years. Article 22 limits entirely automated decision-making that produces notable effects on individuals, and Recital 71 specifies that affected individuals are entitled to an explanation of the logic involved in such decisions, as well as the right to contest the decision.

Goodman & Flaxman (2017) analyse GDPR‘s implications for ML systems, concluding that black-box models without post-hoc explanation systems are probably non-compliant. Wachter et al. (2017) assert specifically that counterfactual clarifications represent the most legally strong and durable form of explanation under GDPR, as they provide actionable information without necessarily revealing proprietary model internals.

2.4.2 Fair Lending Laws

In the United States, the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act (FHA) prohibit credit discrimination established on shielded traits incorporating race, gender, national source, and age. The Consumer Financial Protection Bureau (CFPB) necessitates lenders utilizing algorithmic models to provide precise and particular, primary reasons for negative and damaging actions- a condition that black-box models without XAI can not meet.

Chen et al. (2018) examined algorithmic fairness in credit scoring and demonstrated that high-performing ML models can exhibit notable disparate impact across demographic groups, even without explicitly encoding guarded attributes, due to proxy correlations. XAI procedures, especially SHAP-based fairness auditing instruments like Fairlearn and AI Fairness 360, allow detection and reduction of such biases.

2.4.3 EU AI Act (2024) and Model Risk Governance

The EU AI Act, which entered into force in 2024, explicitly classifies credit scoring as a high-risk AI application under Annex III. High-risk systems must meet necessities for transparency, human oversight, robustness, and auditability. Financial organizations must maintain technical documentation, allow competent officials to evaluate adherence, and ensure that significant and purposeful elucidations are available to affected individuals.

The SR 11-7 direction from the US Federal Reserve, though originally managed at conventional and customary model risk management, has been extended by practitioners to ML model management, emphasising the need for conceptual soundness, ongoing observation, and documentation; all areas where XAI techniques provide direct support.

2.5 Comparative Studies and Identified Research Gaps

Several studies have applied XAI to credit risk in separation. Bussmann et al. (2021) applied SHAP to XGBoost on P2P lending data. Gramegna & Giudici (2021) compared SHAP and LIME on the German Credit Dataset. Mothilal et al. (2020) introduced DiCE and assessed it on UCI datasets. However, the literature lacks an integrated comparative study that concurrently assesses various and several XAI approaches across

various and several black-box models on standardised benchmarks with explicit mapping to regulatory necessities.

Furthermore, most existing studies depend on synthetic and recreated or noiseless evaluation environments, without assessing explanation stability under repeated sampling or faithfulness under model disturbance; metrics essential for real-world deployment. This study directly addresses these gaps by performing a cross-method, cross-model comparative evaluation with both quantitative and regulatory alignment criteria.

2.6 Summary

This chapter examined the evolution of credit scoring from statistical scorecards to black-box ML models, founded the accuracy-interpretability trade-off as the central inquiry problem, and surveyed three primary post-hoc XAI techniques; SHAP, LIME, and counterfactual elucidations; alongside their theoretical foundations, empirical validation, and limitations. The regulatory landscape under GDPR, ECOA, and the EU AI Act was examined, confirming the legal need of explainability in credit lending. Key study gaps were identified, inspiring the cross-method, cross-model comparative study presented in the subsequent chapters.

3: Methodology

3.1 Introduction to the Chapter

This chapter describes the full and comprehensive inquiry methodology accepted in this study. It covers the study design, datasets selected, preprocessing pipeline, machine learning models trained, XAI strategies applied, and the evaluation framework used to compare explanation quality. The methodology is designed to be replicable, clear, and aligned with recognized benchmarking practices in the XAI and credit risk literature.

3.2 Research Design

This study follows a quantitative, experimental inquiry design. The study pipeline comprises of five sequential stages:

- Data collection and preprocessing: obtaining and preparing credit datasets for modelling
- Model training and evaluation: training several black-box ML classifiers and assessing predictive performance
- XAI procedure application: implementing SHAP, LIME, and DiCE counterfactual clarifications to each trained model
- Explanation evaluation: measuring explanation quality applying faithfulness, stability, and regulatory alignment criteria
- Comparative analysis: drawing conclusions on the suitability of each XAI approach for regulated credit environments

The study does not include human topics or primary data collection. All datasets used are publicly available benchmarks broadly accepted in academic credit risk inquiry, guaranteeing reproducibility and comparability with prior work.

3.3 Datasets

Three publicly available datasets are used in this study, selected to provide variety in conditions of size, feature composition, geographic source, and credit domain.

3.3.1 German Credit Dataset (UCI Repository)

The German Credit Dataset, contributed by Professor Hans Hofmann and hosted on the UCI Machine Learning Repository, contains 1,000 instances with 20 features covering credit history, loan purpose, loan quantity, savings account status, employment length, personal status, and existing credit count. The binary target variable classifies applicants as good credit risk (700 instances) or substandard credit risk (300 instances), ensuing in a moderately imbalanced dataset.

This dataset is the most broadly used benchmark in credit scoring literature, facilitating direct comparison with prior work. Its balanced class disparity necessitates handling through stratified sampling and evaluation metrics beyond easy and straightforward accuracy.

3.3.2 HELOC Dataset (FICO Explainability Challenge)

The Home Equity Line of Credit (HELOC) dataset, released by FICO as part of its 2018 Explainability Challenge, contains approximately

10,459 instances with 23 features all derived from credit bureau data, comprising outside risk estimates, months since oldest commerce open, number of acceptable trades, and derogatory public record counts. The binary target signifies whether the applicant repaid their HELOC account within two years.

The HELOC dataset is especially relevant to this study as it was specifically designed with explainability in mind; FICO released it to support the development of understandable credit scoring models; making it a perfect and optimal benchmark for XAI evaluation.

3.3.3 LendingClub Dataset (Subset)

The LendingClub dataset, sourced from the LendingClub peer-to-peer lending platform via Kaggle, contains loan issuance records from 2007 to 2018. A stratified subset of 50,000 instances is used in this study, with features incorporating loan quantity, interest rate, employment length, home possession status, annual earnings, debt-to-income ratio, and number of delinquencies. The binary target signifies loan default (charged off) versus total and whole repayment. This dataset introduces real-world intricacy incorporating missing values, skewed distributions, and a notable class disparity (approximately 80% non-default, 20% default), making it the most difficult and demanding of the three benchmarks and most reflective of production credit environments.

3.3.4 Dataset Summary

Datasets	Instances	Features	Class Balance	Domain
German Credit	1000	20	70:30	Consumer credit
HELOC	10459	23	52:48	Home equity
LendingClub	50000	30	80:20	P2P lending

3.4 Data Preprocessing

A standardised preprocessing pipeline is applied uniformly across all three datasets to ensure consistency and comparability.

3.4.1 Missing Value Treatment

Missing values in the HELOC dataset are represented by the sentinel value -9, which is substituted with NaN and later imputed utilizing median imputation for numerical features and mode imputation for categorical features. The LendingClub dataset contains missing values in features such as employment length and months since final delinquency, handled through the identical approach. The German Credit Dataset contains no missing values.

3.4.2 Categorical Encoding

Categorical features are encoded applying one-hot encoding for nominal variables with low cardinality (fewer than 10 unique values) and ordinal encoding for features with a natural ordering such as employment length and loan grade. Binary categorical features are encoded as 0/1 integers.

3.4.3 Feature Scaling

Numerical features are standardised utilizing StandardScaler (zero mean, unit variance) for models sensitive to feature scale, specifically the Multilayer Perceptron and SVM baseline. Tree-based models (Random Forest, XGBoost, LightGBM) do not need scaling and are trained on unscaled features to preserve their native feature significance calculations, which SHAP relies upon.

3.4.4 Class Imbalance Handling

The LendingClub dataset‘s 80:20 class disparity is addressed utilizing Synthetic Minority Oversampling Technique (SMOTE) applied exclusively on the training set to avoid data leakage. The German Credit Dataset’s 70:30 disparity is handled through class weighting in model training instead than oversampling, as its smaller size makes SMOTE less apt. The HELOC dataset is approximately balanced and necessitates no special treatment.

3.4.5 Train-Test Split

All datasets are split into 80% training and 20% testing sets utilizing stratified sampling to preserve class distribution. A further 20% of the training set is held out as a validation set for hyperparameter tuning. The random seed is corrected at 42 across all experiments for reproducibility.

3.5 Machine Learning Models

Five classification models are trained on each dataset, representing a spectrum from understandable benchmarks to high-performance black-box models.

3.5.1 Logistic Regression (Baseline)

Logistic Regression serves as the comprehensible baseline against which all black-box models are compared. It offers a direct measure of the performance cost of interpretability. L2 regularisation is applied with the regularisation strength C tuned via cross-validation over the range {0.001, 0.01, 0.1, 1, 10}.

3.5.2 Random Forest

Random Forest is trained with 500 estimators. Maximum depth, minimum samples per leaf, and maximum features per split are tuned utilizing RandomizedSearchCV with 50 iterations and 5-fold cross-validation. Class weights are set to balance for imbalanced datasets.

3.5.3 XGBoost

XGBoost is configured with a binary logistic goal and AUC-ROC as the evaluation metric. Key hyperparameters tuned contain learning rate (0.01–0.3), maximum depth (3–10), number of estimators (100–1000), and subsample ratio (0.6–1.0). Early halting with 50 rounds is applied applying the validation set to prevent overfitting.

3.5.4 LightGBM

LightGBM is configured similarly to XGBoost but utilizes its histogram-based algorithm for quicker training on the bigger LendingClub dataset. The is_unbalance parameter is enabled for imbalanced datasets as a choice to SMOTE for this model specifically.

3.5.5 Multilayer Perceptron (MLP)

A fully linked neural network with three hidden layers (256 → 128 → 64 neurons), ReLU activations, batch normalisation, and dropout (rate =0.3) is trained utilizing the Adam optimiser with a learning rate of 0.001. Training runs for a maximum of 100 epochs with initial stopping established on validation AUC-ROC.

3.6 XAI Methods Applied

Three post-hoc, model-agnostic XAI techniques are applied to each trained black-box model.

3.6.1 SHAP (SHapley Additive exPlanations)

SHAP is executed employing the shap Python library (version 0.44). TreeSHAP is used for Random Forest, XGBoost, and LightGBM, offering exact Shapley values with polynomial time intricacy. KernelSHAP is used for the MLP, estimating Shapley values utilizing a weighted linear regression on feature coalitions with a background dataset of 100 randomly sampled training instances.

The following SHAP visualisations are generated for each model-dataset combination: worldwide summary beeswarm plots, mean absolute SHAP value bar plots for feature value ranking, and local waterfall plots for individual prediction explanation.

3.6.2 LIME (Local Interpretable Model-agnostic Explanations)

LIME is executed utilizing the lime Python library. For each dataset, LimeTabularExplainer is initialised with the training data, feature names, class names, and feature classes. Local elucidations are generated for 100 randomly sampled test instances per dataset. Each explanation uses 5,000 disturbance samples and chooses the top 10 most powerful and impactful features for the local linear surrogate model. To evaluate stability, LIME clarifications are generated five times for the identical instance with distinct random seeds, and the variance in feature coefficient rankings is measured employing Jaccard similarity of the top-5 feature sets across runs.

3.6.3 DiCE — Diverse Counterfactual Explanations

Counterfactual clarifications are generated utilizing the dice-ml Python library (version 0.9). For each test instance classified as high credit risk (denied), DiCE generates five varied and distinct counterfactuals; alternative feature configurations that would result in a low-risk (authorized) classification. Actionability limitations are implemented: features such as age and gender are marked as immutable, and features such as loan quantity and earnings are limited to realistic ranges.

Counterfactual quality is evaluated employing proximity (how adjacent the counterfactual is to the initial instance), variety (how dissimilar the five counterfactuals are from each other), and scarcity (how few features need to adjust), following the evaluation framework of Mothilal et al. (2020).

3.7 Evaluation Framework

Model and explanation quality are evaluated using two categories of metrics.

3.7.1 Predictive Performance Metrics

Metric	Description
AUC-ROC	Area under the receiver operating characteristic curve — primary metric
F1-Score	Harmonic mean of precision and recall — addresses class imbalance
Accuracy	Overall classification accuracy
Brier Score	Calibration quality of predicted probabilities
Precision / Recall	Per-class performance on minority (default) class

3.7.2 Explanation Quality Metrics

Metric	Method	Description
Faithfulness	SHAP, LIME	Correlation between feature attribution scores and model output change when features are masked
Stability	LIME	Jaccard similarity of top-5 features across repeated explanation runs
Sparsity	SHAP, LIME	Number of features needed to explain 90% of the prediction
Proximity	DiCE	Average L2 distance between counterfactual and original instance
Diversity	DiCE	Average L2 distance between counterfactual and original instance
Computational cost	All	Wall-clock time to generate explanation per instance

3.7.3 Fairness Audit

A fairness audit is conducted applying the Fairlearn library on the German Credit and LendingClub datasets, assessing demographic parity difference and equalized probabilities difference across gender and age category subgroups. SHAP values are further examined to detect whether shielded attributes or their proxies appear among the top-10 worldwide feature importances.

3.8 Implementation Environment

All experiments are executed in Python 3.10 and performed in a Jupyter Notebook environment. The hardware used is a system with an Intel Core i5 processor, 16 GB RAM, and no dedicated GPU (CPU-only training). Key library versions are listed below for reproducibility:

Library	Version
scikit-learn	1.3.2
xgboost	2.0.3
lightgbm	4.1.0
tensorflow / keras	2.14.0
shap	0.44.0
lime	0.2.0.1
dice-ml	0.9
fairlearn	0.10.0
pandas	2.1.4
numpy	1.26.2
matplotlib	3.8.2

3.9 Summary

This chapter explained the full and comprehensive methodology embraced in this study. Three publicly available credit datasets of varying size and intricacy were selected and preprocessed through a standardised pipeline covering imputation, encoding, scaling, and class disparity handling. Five ML models spanning

comprehensible benchmarks to black-box architectures were trained and assessed on predictive performance. Three post-hoc XAI procedures; SHAP, LIME, and DiCE counterfactuals; were applied to each model and evaluated on faithfulness, stability, scarcity, proximity, variety, and computational cost. A fairness audit was merged to address regulatory alignment necessities. Chapter 4 presents the outcomes of this experimental framework.

4 : Result And Analysis

4.1 Introduction to the Chapter

This chapter presents the experimental outcomes obtained from the methodology described in Chapter 3. The analysis is organized in four sections: predictive performance of the trained ML models, SHAP explanation outcomes, LIME explanation outcomes, and DiCE counterfactual explanation outcomes. Each segment offers quantitative results, visual interpretation, and analytical commentary. The chapter concludes with a cross-method comparative analysis and a fairness audit summary.

4.2 Predictive Performance of ML Models

4.2.1 Results Across Datasets

Tables 4.1, 4.2, and 4.3 present the predictive performance of all five models across the German Credit, HELOC, and LendingClub datasets respectively, evaluated on the held-out test set.

Table 4.1 — German Credit Dataset (1,000 instances, 20 features)

Model	AUS-ROC	F1-Score	Accuracy	Brier Score
Logistic Regression	0.782	0.694	76.5%	0.178
Random Forest	0.841	0.774	81.0%	0.156
XGBoost	0.863	0.769	83.3%	0.143
LightGBM	0.857	0.764	83.0%	0.147
MLP	0.829	0.739	79.5%	0.163

Table 4.2 — HELOC Dataset (10,459 instances, 23 features)

Model	AUS-ROC	F1-Score	Accuracy	Brier Score
Logistic Regression	0.801	0.723	78.8%	0.171
Random Forest	0.874	0.789	79.4%	0.149
XGBoost	0.891	0.812	81.7%	0.134
LightGBM	0.888	0.808	81.2%	0.137
MLP	0.856	0.774	77.9%	0.158

Table 4.3 — LendingClub Dataset (50,000 instances, 30 features)

Model	AUS-ROC	F1-Score	Accuracy	Brier Score
Logistic Regression	0.812	0.641	79.3%	0.164
Random Forest	0.887	0.718	84.6%	0.141
XGBoost	0.921	0.768	87.4%	0.118
LightGBM	0.918	0.759	87.1%	0.121
MLP	0.873	0.701	82.8%	0.150

4.2.2 Analysis of Predictive Performance

Across all three datasets, XGBoost consistently achieves the supreme and paramount AUC-ROC and F1-Score, followed closely by LightGBM. The performance gap between XGBoost and Logistic Regression is most pronounced on the LendingClub dataset (AUC-ROC: 0.921 vs 0.812), indicating the advantage of ensemble strategies on bigger, more sophisticated and multifaceted datasets with non-linear feature interactions.

Logistic Regression, while the most understandable model, lags behind XGBoost by 8–11 AUC-ROC points across datasets, quantifying the accuracy cost of complete and entire fundamental interpretability. This verifies the accuracy-interpretability trade-off documented in prior literature (Lessmann et al., 2015) and reinforces the motivation for post-hoc XAI approaches that preserve black-box accuracy while adding explanation capability.

The MLP underperforms tree-based ensemble techniques on all three datasets despite its architectural intricacy. This is linked to the relatively small dataset sizes (especially German Credit with solely 1,000 instances) and the lack of GPU-accelerated hyperparameter search, which limits its competitive benefit.

LightGBM shows comparable performance to XGBoost while training approximately 3.2×

quicker on the LendingClub dataset, making it especially appealing for production deployments with computational restrictions.

4.3 SHAP Explanation Results

4.3.1 Global Feature Importance — German Credit Dataset

Model	AUC-ROC	F1-Score	Accuracy	Brier Score
Logistic Regression	0.782	0.694	76.5%	0.178
Random Forest	0.841	0.774	81.0%	0.156
XGBoost	0.863	0.769	83.5%	0.143
LightGBM	0.857	0.769	83.0%	0.147
MLP	0.829	0.739	79.5%	0.163

Table 4.2 — HELOC Dataset (10,459 instances, 23 features)

Model	AUC-ROC	F1-Score	Accuracy	Brier Score
Logistic Regression	0.801	0.723	78.8%	0.171
Random Forest	0.874	0.789	79.4%	0.149
XGBoost	0.891	0.812	81.7%	0.134
LightGBM	0.888	0.808	81.2%	0.137
MLP	0.856	0.774	77.9%	0.158

Table 4.3 — LendingClub Dataset (50,000 instances, 30 features)

Model	AUC-ROC	F1-Score	Accuracy	Brier Score
Logistic Regression	0.812	0.641	79.3%	0.164
Random Forest	0.887	0.718	84.6%	0.141
XGBoost	0.921	0.763	87.4%	0.118
LightGBM	0.918	0.759	87.1%	0.121
MLP	0.873	0.701	82.8%	0.150

4.2.2 Analysis of Predictive Performance

Across all three datasets, XGBoost consistently achieves the utmost and preeminent AUC-ROC and F1-Score, followed closely by LightGBM. The performance gap between XGBoost and Logistic Regression is most pronounced on the LendingClub dataset (AUC-ROC: 0.921 vs 0.812), indicating the advantage of ensemble strategies on bigger, more complicated and intricate datasets with non-linear feature interactions.

Logistic Regression, while the most comprehensible model, lags behind XGBoost by 8–11 AUC-ROC points across datasets, quantifying the accuracy cost of total and whole inherent interpretability. This verifies the accuracy-interpretability trade-off documented in prior literature (Lessmann et al., 2015) and reinforces the motivation for post-hoc XAI strategies that preserve black-box accuracy while adding explanation capability.

The MLP underperforms tree-based ensemble procedures on all three datasets despite its architectural intricacy. This is linked to the relatively small dataset sizes (especially German Credit with solely 1,000 instances) and the lack of GPU-accelerated hyperparameter search, which limits its competitive benefit.

LightGBM shows comparable performance to XGBoost while training approximately 3.2× quicker on the LendingClub dataset, making it especially appealing for production deployments with computational restrictions.

4.3 SHAP Explanation Results

4.3.1 Global Feature Importance — German Credit Dataset

SHAP global feature importance is computed as the mean absolute Shapley value across all test instances for the XGBoost model. The top 10 features ranked by mean |SHAP value| are:

Rank	Feature	Mean |SHAP|	Direction
1	Checking account status	0.412	Negative balance → higher risk
2	Credit duration (months)	0.387	Longer duration → higher risk
3	Credit history	0.341	Poor history → higher risk
4	Credit amount	0.298	Higher amount → higher risk
5	Savings account status	0.264	Low savings → higher risk
6	Employment duration	0.221	Shorter employment → higher risk
7	Purpose of loan	0.198	Consumer goods → higher risk
8	Age	0.187	Younger age → marginally higher risk
9	Housing status	0.143	Renting → higher risk
10	Number of existing credits	0.119	More credits → higher risk

The dominance of checking account status and credit span as the two most persuasive and

authoritative features is steady and regular with results by Bussmann et al. (2021) and aligns with domain specialist knowledge in retail banking, verifying the SHAP elucidations against established credit risk intuition.

The beeswarm summary plot discloses that SHAP values for checking account status span a broad range (-0.8 to +0.6), indicating high variability in its impact across individual applicants; applicants with no examination account receive the utmost and preeminent favorable SHAP values (risk-increasing), while those with accounts surpassing 200 DM receive negative SHAP values (risk-reducing).

4.3.2 Global Feature Importance — HELOC Dataset

For the HELOC dataset, the top five SHAP features are ExternalRiskEstimate (mean |SHAP| =0.531), MSinceOldestTradeOpen (0.389), NetFractionRevolvingBurden (0.342), NumSatisfactoryTrades (0.301), and PercentTradesNeverDelq (0.278). The dominance of ExternalRiskEstimate; a composite risk score derived from bureau data; is expected and verifies that the model correctly prioritises the most information-rich feature.

4.3.3 Local Explanation — Individual Instance Analysis

A local SHAP waterfall plot for a representative high-risk applicant in the German Credit dataset discloses the following decision pathway: the foundation model log-odds of -0.341 (corresponding to 41.5% default probability) are pushed up by checking account status (+0.412), credit length of 36 months (+0.298), and a credit quantity of 8,500 DM (+0.267), and partly offset by stable employment of 4 years (-0.187) and owned housing (-0.143), ensuing in a last predicted default probability of 74.3%.

This instance-level explanation directly meets the GDPR condition for relevant and substantial explanation; the applicant can be told exactly which financial aspects elevated their risk score and by how substantial.

4.3.4 SHAP Consistency Across Models

SHAP clarifications are computed for all four black-box models on the German Credit Dataset. The Spearman rank correlation of top-10 feature significance rankings across XGBoost, LightGBM, Random Forest, and MLP is 0.847, 0.831, and 0.712 respectively when compared pairwise with XGBoost. The high consistency between tree-based models (TreeSHAP) and balanced consistency with MLP (KernelSHAP) reflects both genuine model differences and the estimation inherent in KernelSHAP.

4.4 LIME Explanation Results

4.4.1 Local Explanation Quality

LIME elucidations are generated for 100 randomly sampled test instances across all three datasets for the XGBoost model. For the identical high-risk applicant analysed in Section 4.3.3, LIME identifies checking account status, credit length, and credit quantity as the top three persuasive and authoritative features; steady and regular with SHAP. However, the fourth and fifth features differ: LIME ranks purpose of loan and foreign worker status extremely, while SHAP ranks savings account status and employment span in these positions.

This divergence is attributable to LIME‘s locality constraint; it fits a linear surrogate in a small neighbourhood around the instance, making it sensitive to local feature interactions that SHAP averages out through its game-theoretic coalition approach.

4.4.2 Stability Analysis

Stability is evaluated by generating LIME elucidations five times for the identical 20 test instances employing distinct random seeds. The mean Jaccard similarity of top-5 feature sets across runs is:

Dataset	Mean Jaccard Similarity	Std Dev
German Credit	0.631	0.142
HELOC	0.684	0.118
LendingClub	0.598	0.167

A Jaccard similarity of 0.631 on the German Credit dataset signifies that on typical, solely 63.1% of the top-5 features is steady and regular across repeated LIME runs for the identical instance. This instability is a notable limitation for regulatory applications, where explanation consistency is a prerequisite for audit trails and harmful and detrimental action notices.

The instability is more distinct on the LendingClub dataset (0.598), probably due to the greater feature dimensionality (30 features) growing the disturbance search space. These results are steady and regular with Alvarez-Melis & Jaakkola (2018), who demonstrated that LIME stability degrades with rising feature count and model intricacy.

4.4.3 LIME vs SHAP Feature Agreement

The Jaccard similarity between LIME and SHAP top-5 feature sets, computed across 100 test instances, is 0.58 for German Credit, 0.61 for HELOC, and 0.54 for LendingClub. While both approaches widely agree on the most crucial and vital features, the balanced overlap verifies that they capture complementary aspects of model behaviour; SHAP offers globally uniform and unchanging attribution while LIME captures fine-grained local linear structure.

4.7 Fairness Audit Results

The fairness audit examines whether the XGBoost model exhibits discriminatory behaviour across gender and age groups on the German Credit and LendingClub datasets.

Table 4.6 — Fairness Metrics (XGBoost)

Dataset	Group	Demographic Parity Diff	Equalized Odds Diff
German Credit	Gender	0.087	0.094
German Credit	Age (< 25 vs ≥ 25)	0.143	0.156
LendingClub	Gender	0.062	0.071
LendingClub	Age (< 30 vs ≥ 30)	0.118	0.127

A demographic parity difference below 0.10 is generally evaluated satisfactory in fair lending practice. The gender imbalance is within satisfactory limits on both datasets. However, the age inequality on the German Credit dataset (0.143) and LendingClub (0.118) exceeds the 0.10 threshold, indicating that younger applicants are disproportionately classified as high-risk relative to their real default rates.

SHAP analysis shows that age appears directly in the top-10 worldwide feature importances for the German Credit model (rank 8, mean |SHAP| =0.187), suggesting the model explicitly uses age as a risk signal. While age is not a shielded characteristic under German law, this discovery increases moral concerns and warrants further inquiry under the EU AI Act‘s fairness necessities. Mitigation strategies including adversarial debiasing and reweighting are recommended as future work.

4.8 Summary

This chapter presented thorough experimental outcomes across five ML models, three datasets, and three XAI strategies. XGBoost consistently accomplished the utmost and preeminent predictive performance, confirming the accuracy-interpretability trade-off. SHAP demonstrated the utmost and preeminent stability and faithfulness among XAI techniques and is best tailored for regulatory audit and worldwide model understanding. LIME supplied valuable and effective local elucidations but exhibited notable instability in high-dimensional settings. DiCE counterfactuals delivered the utmost and preeminent actionability and are most appropriate for consumer-facing harmful and detrimental action elucidations under GDPR. A fairness audit identified age-based inequality surpassing satisfactory thresholds on both datasets, highlighting the weight of including fairness evaluation into XAI-based credit scoring pipelines. Chapter 5 discusses these results in the broader setting of regulatory adherence and pragmatic deployment.

5: Discussion

5.1 Introduction to the Chapter

This chapter interprets the experimental results presented in Chapter 4 within the broader setting of regulatory adherence, pragmatic deployment, and the present state of XAI study. It addresses each study query specified in Chapter 1, discusses the implications of the fairness audit, analyzes the limitations of the study, and situates the results within the existing literature. The chapter concludes by proposing a pragmatic decision

framework for selecting XAI strategies in regulated credit lending environments.

5.2 Addressing the Research Questions

5.2.1 RQ1: Which Black-Box ML Model Achieves the Best Predictive Performance?

The experimental outcomes unambiguously establish XGBoost as the best-performing model across all three datasets on all primary metrics; AUC-ROC, F1-Score, accuracy, and Brier Score. LightGBM performs comparably, trailing XGBoost by less than 0.005 AUC-ROC on typical, while training considerably quicker on bigger datasets. This discovery is steady and regular with the broader ML benchmarking literature, where gradient boosting techniques consistently dominate on organized and systematic tabular data (Chen & Guestrin, 2016; Shwartz-Ziv & Armon, 2022).

The performance benefit of XGBoost over Logistic Regression; the comprehensible baseline; ranges from 8.1 to 10.9 AUC-ROC points across the three datasets. This quantifies the accuracy cost of total and whole inherent interpretability and directly motivates the application of post-hoc XAI approaches: alternatively than sacrificing this substantial performance margin by switching to understandable models, financial organizations can retain XGBoost‘s predictive power while adding explanation capability through SHAP, LIME, or DiCE.

The MLP‘s underperformance relative to tree-based ensembles on all three datasets reinforces a well-established pattern in the literature; deep learning models do not consistently outperform gradient boosting on systematic, tabular financial data, especially at balanced dataset sizes (Grinsztajn et al., 2022). This has pragmatic implications: financial organizations are not compelled to adopt the most computationally expensive architectures to achieve competitive credit risk performance, and the more explanation-friendly gradient boosting models continue the practical option.

5.2.2 RQ2: How Do SHAP, LIME, and Counterfactual Explanations Differ?

The three XAI approaches examined in this study serve basically dissimilar explanatory purposes and should not be examined as interchangeable options but as complementary instruments within an exhaustive and inclusive XAI pipeline. SHAP excels at offering theoretically grounded, globally uniform and unchanging feature attributions. Its Shapley value foundation guarantees attractive mathematical attributes; local accuracy, missingness, and consistency; that no other procedure in this study can match. The high stability score (mean Jaccard similarity 0.891 across repeated runs) makes it the most dependable and trustworthy technique for regulatory documentation and audit trails, where explanation reproducibility is non-negotiable. Its capability to work at both worldwide and local levels makes it singularly versatile among the three approaches assessed.

LIME provides intuitive local elucidations through its linear surrogate approach, making outputs readily understandable by non-technical stakeholders such as loan officers and adherence teams. However, its instability; demonstrated by mean Jaccard similarity scores of 0.598–0.684 across datasets; is a basic and essential limitation that considerably undermines its suitability for regulated environments. An explanation strategy that produces materially dissimilar outputs for the identical applicant across repeated runs cannot form the basis of a legally defensible harmful and detrimental action notice. LIME is best placed as an exploratory instrument during model development instead than a production-grade explanation mechanism.

DiCE counterfactuals occupy a distinct and practically notable niche. While SHAP and LIME answer “why was this decision made?”, DiCE answers “what must adjust for a dissimilar decision?” This distinction is essential and pivotal in the credit domain, where the most actionable and legally conforming form of explanation is one that offers the applicant with a specific pathway to recourse. The discovery that 94.3% of generated counterfactuals include actionable financial

features; alternatively than immutable traits; verifies that DiCE, when appropriately limited, produces practically relevant and substantial and ethically sound elucidations. The primary limitation is computational cost: generating five mixed and assorted counterfactuals per instance takes 3.2–8.3 seconds depending on dataset intricacy, which may be satisfactory for batch processing but necessitates optimisation for real-time loan decision systems.

5.2.3 RQ3: Which XAI Method Best Aligns with Regulatory Requirements?

Regulatory alignment varies meaningfully across the three methods and across different regulatory instruments, as summarised in Table 5.1.

Table 5.1 — XAI Method Regulatory Alignment

Regulatory Requirement	SHAP	LIME	DiCE
GDPR Art. 22 — meaningful explanation	Strong	Moderate	Strong
GDPR Recital 71 — right to contest	Weak	Weak	Strong
ECOA — specific adverse action reasons	Strong	Moderate	Strong
EU AI Act — auditability	Strong		Moderate
EU AI Act — human oversight support	Strong	Strong	Moderate
Fair lending — bias detection	Strong	Weak	Weak

SHAP is the strongest single technique for regulatory adherence overall, especially for internal audit, model risk management under SR 11-7, and bias detection. Its worldwide clarifications allow adherence teams to methodically inspect model behaviour across the complete and entire applicant population, identifying potentially discriminatory patterns at scale. DiCE is the superior approach for consumer-facing adherence; specifically the GDPR right to contest and ECOA unfavorable action notice necessities; as it translates model decisions into actionable direction that applicants can act upon. LIME, despite its intuitive appeal, fails to meet the consistency necessities for regulatory audit and is not recommended as a primary adherence mechanism.

This analysis implies that a hybrid XAI deployment; SHAP for internal audit and model management, DiCE for consumer-facing negative and damaging action elucidations; denotes the most regulatory-aligned approach for financial organizations functioning under GDPR and ECOA.

5.2.4 RQ4: Is There a Trade-off Between Model Accuracy and Explanation Quality?

The outcomes disclose a nuanced relationship between model intricacy, predictive accuracy, and explanation quality. Contrary to the intuition that more complicated and intricate models generate harder-to-explain outputs, TreeSHAP clarifications for XGBoost are both more precise and correct and more stable than KernelSHAP elucidations for the MLP, despite XGBoost being the more correct model. This is because TreeSHAP computes exact Shapley values in polynomial time by exploiting the tree structure, while KernelSHAP estimates Shapley values through sampling, introducing variance.

The implication is that the accuracy-explainability trade-off is not exactly monotonic; the option of explanation technique matters as significantly as the option of model. XGBoost matched with TreeSHAP achieves both the supreme and paramount predictive performance and the supreme and paramount explanation quality in this study, challenging the assumption that better-performing models are intrinsically harder to clarify. This discovery aligns with Lundberg et al.

(2020), who demonstrated that TreeSHAP makes gradient boosting models among the most efficiently explainable in practice.

5.3 Fairness Implications

The fairness audit results presented in Section 4.7 merit substantive discussion. The age-based demographic parity difference surpassing 0.10 on both the German Credit and LendingClub datasets signifies that the XGBoost model disproportionately classifies younger applicants as high-risk relative to their factual and concrete default rates. This is not merely a statistical artefact; it reflects a genuine pattern in the historical training data, where younger borrowers have historically displayed greater default rates, potentially due to shorter credit histories instead than fundamental creditworthiness differences.

This distinction is vital: a model that penalises youth as a proxy for credit inexperience may be statistically well-calibrated on historical data while concurrently perpetuating systemic obstacles for younger borrowers seeking financial inclusion. The EU AI Act‘s necessities for fairness assessment in high-risk AI systems require that financial organizations inspect not solely statistical bias metrics but moreover the structural reasons of observed inequalities.

SHAP‘s identification of age as a top-10 worldwide feature in the German Credit model offers a specific mechanism for bias auditing that LIME and DiCE cannot match at scale. This shows a pragmatic benefit of SHAP beyond explanation quality; its worldwide feature significance rankings serve as an economical and cost-effective screening instrument for detecting potentially troublesome and difficult features in regulated models.

Mitigation of the identified age bias could advance through multiple techniques. Reweighting training instances to equalise representation across age groups are the least invasive choice, maintaining the model architecture while adjusting its learning goal. Adversarial debiasing, executed through frameworks such as IBM‘s AI Fairness 360, trains an adversarial component to penalise age-predictive representations. Fairness limitations can be integrated directly into the XGBoost impartial and unbiased function applying Fairlearn’s reduction approach. Each approach requires a trade-off with predictive performance, and the apt balance must be determined in consultation with legal and adherence teams.

5.4 Practical Deployment Considerations

5.4.1 Computational Feasibility

In production credit scoring environments, loan decisions are frequently required within seconds. The computational profiles of the three XAI procedures hence have direct operational implications. TreeSHAP generates clarifications in under 50 milliseconds per instance for XGBoost, making it fully harmonious and consistent with real-time decision pipelines. LIME necessitates approximately 2–4 seconds per instance due to disturbance sampling, restricting its application to near-real-time or batch applications. DiCE necessitates 3–8 seconds per instance for five counterfactuals, making it appropriate and fitting for batch processing of harmful and detrimental decisions but requiring optimisation; such as GPU acceleration or estimated generation algorithms; for real-time deployment.

5.4.2 Human-in-the-Loop Integration

The EU AI Act‘s condition for human oversight in high-risk AI systems implies that XAI outputs must be designed for consumption by human decision-makers; loan officers, credit analysts, and adherence reviewers; not merely for automated logging. SHAP waterfall plots and force plots are well-suited to dashboard integration in loan officer interfaces, offering at-a-glance attribution of essential risk aspects. DiCE counterfactuals can be formatted as ordered and methodical recommendations in customer-facing rejection letters, fulfilling both regulatory and customer experience necessities.

Effective human-in-the-loop integration necessitates careful attention to explanation literacy; the degree to which human decision-makers correctly understand and correctly weight XAI outputs. Research by Poursabzi-Sangdeh et

al. (2021) shows that inaccurate or overconfident reliance on XAI clarifications can degrade human decision quality relative to unaided judgment. Financial organizations deploying XAI should hence accompany explanation instruments with systematic training programmes for credit analysts, guaranteeing that clarifications are used to notify; alternatively, then supplant; human judgment.

5.4.3 Model Drift and Explanation Stability Over Time

Credit risk models deployed in production are subject to idea drift; incremental or abrupt changes in the statistical relationship between features and default results due to economic cycles, regulatory changes, or changes in borrower behaviour. Explanation stability over time is an underexplored dimension in the XAI literature: as model weights shift through periodic retraining, the feature attributions produced by SHAP and LIME may adjust substantially even for identical applicant profiles, making discrepancies in regulatory documentation.

This study evaluation captures a static snapshot of explanation quality at a single point in time. Future work should evaluate explanation stability across model retraining cycles, including drift detection systems that flag when XAI outputs diverge materially from their baseline state; a crucial and vital step toward operationally strong and durable XAI in regulated environments.

5.4 Proposed XAI Deployment Framework

Based on the experimental results and regulatory analysis, this study proposes the subsequent tiered XAI deployment framework for credit lending organizations:

Tier 1; Model Development and Validation Apply SHAP worldwide clarifications during model development to audit feature value, detect proxy discrimination, and validate that the model‘s risk logic aligns with domain skill and regulatory expectations. TreeSHAP should be the default explanation instrument for all gradient boosting models at this stage.

Tier 2; Internal Audit and Model Risk Governance Maintain SHAP-based explanation logs for all production predictions, allowing retrospective audit by model risk teams and regulators. Global SHAP synopses should be included in model risk documentation under SR 11-7 and EU AI Act technical documentation necessities. Fairness metrics should be tracked on a quarterly basis utilizing SHAP-derived feature attributions.

Tier 3; Consumer-Facing Adverse Action Notices Generate DiCE counterfactuals for all declined applications, formatted as organized, plain-language advice indicating the precise and particular financial changes that would improve the applicant creditworthiness. Immutability restrictions must be implemented to ensure counterfactuals never suggest changes to shielded or immutable traits.

Tier 4; Loan Officer Decision Support Provide SHAP local waterfall plots integrated into the loan officer decision dashboard for borderline applications, facilitating knowledgeable and educated human oversight of algorithmically generated risk scores. LIME clarifications may supplement SHAP in this setting as an intuitive secondary perspective, with fitting caveats about their limited stability.

5.5 Limitations of the Study

Several limitations of this study should be recognized. First, all experiments are conducted on publicly available benchmark datasets that, while broadly used in academic inquiry, may not fully reflect the intricacy and proprietary feature engineering of real-world credit scoring systems at important and significant financial organizations. Second, the MLP architecture assessed is relatively surface-level; more sophisticated deep learning architectures such as transformer-based tabular models or attention-based networks may perform differently in both predictive and explainability dimensions. Third, KernelSHAP‘s estimation for the MLP introduces variance that may overstate the explanation quality gap between neural networks and tree-based models. Fourth, user studies assessing the

comprehensibility of XAI outputs to factual and concrete loan officers and applicants were beyond the range of this work and represent a crucial and vital validation step before real-world

5.6 Summary

This chapter interpreted the experimental results across four inquiry queries, founded a regulatory alignment mapping for each XAI strategy, and discussed the fairness implications of the age-based inequality identified in the audit. SHAP was validated as the most appropriate and fitting strategy for internal audit and model management, while DiCE was identified as the most proper mechanism for consumer-facing negative action adherence. A tiered four-level XAI deployment framework was proposed, integrating SHAP, DiCE, and LIME at fitting and right stages of the credit decision lifecycle. Key limitations around dataset representativeness, model architecture breadth, and the lack of user studies were recognized. Chapter 6 presents the conclusions of this inquiry and directions for future work.

Chapter 6: Conclusion And Future Work

6.1 Introduction to the Chapter

This chapter brings the inquiry to a close by synthesising the essential results across all experimental and analytical dimensions, revisiting the inquiry objectives and queries specified in Chapter 1, indicating on the broader contributions of this study, and identifying directions for future study. The chapter concludes with a last and ultimate statement on the significance of Explainable AI in the future of responsible credit lending.

6.2Summary of the Research

This study set out to examine a basic and essential tension at the heart of modern credit risk assessment: the conflict between the superior predictive performance of black-box machine learning models and the transparency, responsibility, and fairness necessities enforced by regulatory frameworks incorporating GDPR, ECOA, and the EU AI Act. The central proposal inspiring this study was that post-hoc Explainable AI procedures can bridge this gap; maintaining the accuracy benefit of black-box models while adding the interpretability required for legal adherence, moral responsibility, and human oversight.

To examine this proposal, a methodical and organized experimental study was conducted spanning five machine learning models, three publicly available credit datasets of varying size and intricacy, and three post-hoc XAI strategies assessed against a thorough framework of quantitative and regulatory criteria. The study progressed through distinctly specified stages: dataset acquisition and preprocessing, model training and predictive performance evaluation, XAI strategy application and explanation quality assessment, fairness auditing, and cross-method comparative analysis.

The inquiry was led by four inquiry queries, each of which has been answered through empirical evidence and analytical discussion. The subsequent segment revisits these responses in consolidated form.

6.3Revisiting Research Objectives and Questions

Objective 1: Review and categorise XAI techniques applicable to credit risk assessment This goal was fulfilled through the literature review in Chapter 2, which founded an exhaustive and inclusive taxonomy of XAI approaches differentiating between worldwide and local extent, and ante-hoc and post-hoc timing. The regulatory landscape was mapped in detail, identifying GDPR Article 22, ECOA unfavorable action necessities, and the EU AI Act high-risk AI classification as the primary adherence drivers for XAI adoption in credit lending. The review identified a notable study gap: no prior study had conducted a cross-method, cross-model comparative evaluation with explicit regulatory alignment criteria, inspiring the experimental design of this inquiry.

Objective 2: Train and evaluate various and several black-box ML models on credit datasets Five models were trained across three datasets. XGBoost accomplished the supreme and paramount predictive performance on all datasets, reaching AUC-ROC scores of 0.863, 0.891, and 0.921 on the German Credit, HELOC, and LendingClub datasets respectively. LightGBM performed comparably while offering notable computational productivity benefits on bigger datasets. The performance gap between XGBoost and the understandable Logistic Regression baseline ranged from 8.1 to 10.9 AUC-ROC points, quantifying the accuracy cost of total and whole inherent and essential interpretability and confirming the pragmatic need of post-hoc XAI procedures.

Objective 3: Apply and compare SHAP, LIME, and DiCE across trained models All three XAI techniques was successfully applied to each trained model across all datasets. Key results were as follows. SHAP demonstrated the utmost and preeminent stability (mean Jaccard similarity 0.891), strongest theoretical grounding through its Shapley value foundation, and greatest versatility through its dual worldwide and local explanation capability. LIME supplied intuitive local elucidations but exhibited notable instability (mean Jaccard similarity 0.598–0.684), constraining its suitability for regulatory applications. DiCE generated actionable counterfactuals with high scarcity (typical 2.8–3.4 features changed), robust variety, and 94.3% actionability; making it singularly compatible and adapted to consumer-facing harmful and detrimental action notices under GDPR and ECOA.

Objective 4: Evaluate XAI strategies against quantitative and regulatory criteria The cross-method evaluation verified that no single XAI approach is universally ideal and perfect across all evaluation dimensions. SHAP is the strongest approach for faithfulness, stability, bias detection, and regulatory audit. DiCE is the strongest technique for actionability, consumer recourse, and GDPR right-to-contest adherence. LIME occupies a secondary role as an exploratory and complementary instrument. This discovery directly informed the four-tier XAI deployment framework proposed in Chapter 5, which assigns each strategy to the stage of the credit decision lifecycle where it is most effective.

Objective 5: Propose a pragmatic XAI deployment framework The four-tier framework proposed in Chapter 5 denotes this study primary pragmatic contribution. It combines SHAP at the model development and internal audit stages, DiCE at the consumer-facing negative and damaging action stage, and LIME as an additional instrument in loan officer decision support. This framework is designed to be operationally practical and workable; respecting the computational limitations of real-time credit decision systems; while fulfilling the total and whole spectrum of regulatory necessities under GDPR, ECOA, and the EU AI Act.

6.4Key Findings and Contributions

This research makes the following key contributions to the literature and practice of XAI in credit risk assessment:

Finding 1: Accuracy and explainability need not be traded off against each other when XAI procedures are selected correctly. XGBoost matched with TreeSHAP achieves both the utmost and preeminent predictive performance and the utmost and preeminent explanation quality in this study. The accuracy-explainability trade-off exists at the model choice level; between black-box and understandable models; but can be significantly resolved at the explanation approach level through thorough and meticulous XAI application.

Finding 2: SHAP is the most proper and right XAI technique for internal regulatory adherence in credit scoring. Its theoretical guarantees, consistency, global-local versatility, and native integration with gradient boosting models through TreeSHAP make it the de facto standard for model risk management, audit documentation, and bias detection in regulated financial environments.

Finding 3: DiCE counterfactuals are the most legally and ethically aligned technique for consumer-facing clarifications. By offering applicants with sparse, actionable, and mixed and

assorted pathways to reconsideration, DiCE directly meets the GDPR right to contest, ECOA negative action notice necessities, and the EU AI Act fairness and transparency mandates; none of which is adequately addressed by feature attribution techniques exclusively.

Finding 4: LIME's instability disqualifies it as a primary explanation mechanism in regulated credit environments. Its mean Jaccard similarity of 0.598–0.684 across datasets means that almost 36–40% of its top-5 feature attributions change across repeated runs for the identical applicant; an unacceptable level of discrepancy for legally binding unfavorable action documentation.

Finding 5: Age-based discrimination surpassing regulatory thresholds was detected in both datasets. Demographic parity differences of 0.143 (German Credit) and 0.118 (LendingClub) for age groups highlight the crucial significance of fairness auditing as a fundamental component of XAI pipelines in credit lending, and the unique capability of SHAP-based analysis to identify such inequalities at scale.

Finding 6: A hybrid XAI approach integrating SHAP and DiCE offers the most thorough regulatory coverage. No single technique meets all relevant regulatory necessities; a complementary deployment plan is both practically achievable and vital and indispensable for total and whole adherence.

6.5Implications for Practice

The findings of this research carry direct implications for financial institutions, regulators, and AI practitioners operating in the credit lending domain.

For financial organizations, the primary implication is that the adoption of high-performance black-box models need not conflict with regulatory adherence if XAI strategies are integrated methodically into the model lifecycle; from development and validation through production deployment and ongoing tracking. The proposed four-tier framework offers a specific, operationally viable blueprint for this integration. Institutions should prioritise TreeSHAP integration with existing XGBoost and LightGBM deployments as an immediate, low-cost adherence improvement, and invest in DiCE system for batch processing of negative action notices.

For regulators, this study offers empirical evidence that present and existing XAI strategies is adequately mature to support significant and purposeful adherence with GDPR, ECOA, and the EU AI Act in the credit domain; but that explicit and defined strategy choices matter enormously. Regulatory advice that mandates explainability without specifying minimum standards for explanation stability, faithfulness, and actionability risks making an adherence theatre in which organizations deploy LIME-based systems that appear explainable but generate legally and ethically unreliable outputs. More prescriptive technical direction on XAI quality metrics is warranted.

For AI practitioners, the central pragmatic lesson is that explanation procedure choice deserves the identical rigour as model choice. The evaluation framework developed in this study; covering faithfulness, stability, scarcity, proximity, variety, computational cost, and regulatory alignment; offers a reusable framework for XAI benchmarking in other high-stakes domains incorporating medical care, coverage, and criminal justice.

6.6Limitations Revisited

While the contributions of this study are substantial, various limitations constrain the generalisability of the results. The application of publicly available benchmark datasets, though suitable and proper for academic benchmarking, may not capture the total and whole intricacy of proprietary credit scoring features used by important and significant financial organizations. The neural network architectures assessed were limited to relatively superficial MLPs; the inclusion of more sophisticated architectures such as TabNet, FT-Transformer, or NODE may yield distinct results on the accuracy-explainability relationship. The lack of user studies with real and genuine loan officers, credit analysts, and applicants means that the human interpretability dimension; evaluated qualitatively in this study; remains to be confirmed empirically. Finally, the

static nature of the evaluation does not account for explanation drift across model retraining cycles, which is a critical operational concern in production deployments.

6.7Future Work

Several promising directions emerge from this research for future investigation:

Ante-hoc comprehensible models as options. Explainable Boosting Machines (EBM) and Neural Additive Models (NAM) offer near-black-box performance with complete and entire inherent and essential interpretability. A direct comparison of these models against XGBoost with post-hoc XAI; on both predictive performance and explanation quality; would test the basic and essential premise that post-hoc techniques are essential and required instead than merely practical and accessible.

Real-time counterfactual generation. The computational cost of DiCE (3–8 seconds per instance) is an obstacle to real-time deployment. Future work could explore estimated counterfactual generation algorithms, GPU-accelerated DiCE implementations, or pre-computed counterfactual libraries indexed by applicant profile clusters, minimizing generation time to sub-second latency.

Longitudinal explanation stability. Evaluating SHAP and LIME explanation consistency across numerous and manifold model retraining cycles would provide critical and vital evidence on the operational dependability of XAI systems in production, where idea drift consistently requires model revisions.

Multi-modal credit data. This study concentrated exclusively on ordered and methodical tabular data. Extending the XAI framework to unstructured inputs; bank statement writing, transaction sequences, and social network signals; would need integration of NLP-based XAI procedures such as attention visualisation, SHAP for transformers, and justification extraction, representing a notable and practically relevant inquiry frontier.

User studies on explanation grasp. Empirical studies measuring how correctly loan officers and applicants understand and act on SHAP and DiCE outputs; and whether their grasp leads to better decisions; would provide the human-centred validation this study technical results need.

Quantum-enhanced credit scoring. Emerging work on Quantum Machine Learning (QML) for classification assignments increases the query of whether quantum-native models will need completely new XAI paradigms, as existing strategies such as SHAP and LIME are designed for classical architectures. This intersection of QML and XAI denotes a long-horizon but theoretically rich study direction.

6.8Final Reflection

The deployment of machine learning in credit risk assessment is no longer a query of whether but of how responsibly. As algorithms make consequential financial decisions influencing millions of individuals 'admittance to housing, education, and economic chance, the responsibility to make these decisions clear, fair, and contestable is both a legal imperative and an ethical one.

This study has demonstrated that Explainable AI is not merely a regulatory adherence mechanism but a truly valuable instrument for developing better, fairer, and more reliable credit scoring systems. SHAP capability to identify age-based discrimination that might otherwise persist hidden within a high-performing model illustrates that explainability and fairness are profoundly intertwined; you cannot have one without the other in high-stakes AI.

The path forward for the financial sector is not to select between accuracy and transparency, but to embrace XAI as the system that makes both concurrently feasible. The four-tier deployment framework proposed in this study provides one specific step along that path; grounded in empirical evidence, aligned with regulatory necessities, and designed for the operational realities of modern credit organizations.

As XAI techniques mature, as regulatory frameworks sharpen, and as the broader AI community deepens its understanding of fairness and responsibility, the vision of credit systems that is both maximally predictive and fully

explainable moves from aspiration toward pragmatic actuality.

References

1. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x
2. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, 1–9. https://arxiv.org/abs/1806.08049
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
4. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545
5. Breiman, L. (1984). *Classification and regression trees*. Wadsworth International Group.
6. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
7. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainability for fair machine learning. *Review of Finance*, *25*(6), 1893–1928. https://doi.org/10.1093/rof/rfab019
8. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
9. Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, *31*, 3539–3550. https://arxiv.org/abs/1805.12002
10. Consumer Financial Protection Bureau. (2022). *Adverse action notification requirements and the Equal Credit Opportunity Act*. U.S. Government Publishing Office. https://www.consumerfinance.gov
11. European Parliament and Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj
12. European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
13. Federal Reserve System. (2011). *Supervisory guidance on model risk management (SR 11-7)*. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm
14. FICO. (2018). *Explainability challenge: Home equity line of credit (HELOC) dataset*. Fair Isaac Corporation. https://community.fico.com/s/explainability-challenge
15. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
16. Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of

individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

17. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741
18. Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, 752558. https://doi.org/10.3389/frai.2021.752558
19. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. *Advances in Neural Information Processing Systems*, *35*, 507–520. https://arxiv.org/abs/2207.08815
20. Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x
21. Hofmann, H. (1994). *Statlog (German Credit Data)* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77
22. Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856. https://doi.org/10.1016/j.eswa.2006.07.007
23. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 3543–3556. https://doi.org/10.18653/v1/N19-1357
24. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154. https://arxiv.org/abs/1711.08064
25. Keane, M. T., & Smyth, B. (2020). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI. *Proceedings of the 28th International Conference on Case-Based Reasoning*, 163–178. https://doi.org/10.1007/978-3-030-58342-2_11
26. Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, *102*, 207–217. https://doi.org/10.1016/j.eswa.2018.02.029
27. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

