INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

IMPROVING MODEL INTERPRETABILITY USING HYBRID EXPLAINABLE AI TECHNIQUES: A COMBINED SHAP-LIME FRAMEWORK WITH CONFIDENCE SCORING AND STABILITY ANALYSIS

Krishnamoorthy V, Vinaya B
Faculty, Student
Department of Computer Science and Engineering
Sri Venkateswara College of Engineering, Sriperumbudur, India

***Abstract*:** The widespread adoption of machine learning (ML) models in critical applications—healthcare, finance, and criminal justice—demands greater model interpretability. Despite advances in explainability techniques, existing methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have inherent limitations when applied independently. This paper presents a novel Hybrid Explainable AI (XAI) System that combines SHAP and LIME through weighted fusion, complemented by confidence scoring mechanisms and stability analysis. Our framework addresses the fundamental challenge of explanation inconsistency by computing agreement metrics between different explanation methods, enabling practitioners to quantify explanation reliability. We employ Spearman rank correlation and cosine similarity to assess consistency between explanations, producing a unified confidence score that reflects both model prediction confidence and explanation agreement. The proposed system is evaluated on the UCI Diabetes dataset using a RandomForest classifier, achieving 85.2% accuracy while providing interpretable feature importance scores. Results demonstrate that our hybrid approach yields more robust and stable explanations compared to individual methods, with a consistency score of 0.76 (Spearman) between SHAP and LIME. This work contributes a practical framework for enhancing trust in ML systems through reliable, multi-method explanations suitable for regulated industries.

Index Terms - Explainable AI, SHAP, LIME, Hybrid Fusion, Confidence Scoring, Model Interpretability, Hybrid Explanations, Stability Analysis

I. INTRODUCTION

A. Background and Motivation

Machine learning models have become indispensable tools across diverse domains—medical diagnosis, financial risk assessment, autonomous systems, and criminal justice. However, as model complexity increases, particularly with ensemble methods and deep neural networks, the ability to understand and interpret model decisions diminishes. This phenomenon, known as the "black-box problem," poses significant risks in high-stakes applications where decisions directly impact human welfare [1].

The consequences of unexplainable AI systems extend far beyond technical concerns. Consider a scenario where a credit scoring algorithm denies a loan application. Traditional ML systems provide only a numerical prediction (e.g., risk score: 0.72) without explaining which factors influenced the decision. In regulated industries, this lack of transparency violates principles of algorithmic fairness and regulatory compliance, including GDPR Article 22, Fair Lending Regulations, and the algorithmic accountability requirements emerging across jurisdictions [2]. Medical practitioners face similar challenges: clinicians require not just predictions but rigorous explanations to validate diagnostic reasoning against clinical knowledge.

The importance of explainability extends well beyond regulatory compliance. Explainability fosters trust in automated systems—users who understand model reasoning are more likely to accept and appropriately rely upon AI-assisted decisions. Explanations serve as debugging tools, revealing systematic model errors, dataset biases, and spurious feature correlations. Interpreting feature importance guides data collection

strategies and feature engineering efforts. Comparing explanations across model architectures informs model selection. Finally, systematic analysis of explanation patterns enables fairness auditing, identifying features that may encode or amplify discriminatory signals.

B. Limitations of Existing Approaches

Two prominent XAI methods have emerged as foundational approaches: SHAP (SHapley Additive exPlanations) grounded in cooperative game theory, and LIME (Local Interpretable Model-agnostic Explanations) based on local surrogate approximation. While both provide valuable local explanations, each method exhibits significant limitations.

SHAP, while theoretically elegant and providing principled Shapley value-based explanations satisfying desirable axioms (efficiency, symmetry, dummy, additivity), faces severe computational constraints. Exact SHAP computation requires O(2^n) model evaluations in the general case. Additionally, SHAP's conditional expectation formulation implicitly assumes feature independence—an assumption frequently violated in real data. The method also exhibits sensitivity to background dataset selection.

Conversely, LIME offers model-agnosticism and computational efficiency through local surrogate fitting based on perturbation sampling. However, LIME demonstrates high instability across multiple runs: executing LIME twice on identical instances typically yields different feature rankings due to stochastic perturbation sampling. The perturbation strategy significantly influences explanation quality, yet principled guidance for parameter selection remains limited. Furthermore, LIME's local neighborhood constraint prevents establishing global interpretability patterns.

C. Research Gap and Contribution

The fundamental contradiction underlying current XAI practice is clear: neither SHAP nor LIME alone provides sufficiently reliable explanations for deployment in high-stakes applications. More critically, when these independent explanation methods disagree—which occurs frequently in practice—practitioners lack a principled mechanism to assess explanation reliability. Existing literature addresses explanation generation but neglects explanation validation. This critical gap between generating multiple explanations and trusting them represents the primary motivation for our research: we seek to develop a framework that not only generates diverse explanations but also quantifies explanation confidence through principled inter-method agreement analysis.

D. Paper Contributions

This paper makes the following contributions:

(1) Hybrid XAI Framework: A principled approach combining SHAP and LIME through weighted fusion with configurable weights, balancing computational cost against explanation richness.

(2) Confidence Scoring Mechanism: A novel confidence score based on agreement metrics (Spearman correlation, cosine similarity) between explanation methods, quantifying explanation reliability independently of model prediction confidence.

(3) Stability Analysis: Systematic evaluation of explanation consistency across multiple runs, identifying inherent instabilities in individual methods and demonstrating hybrid approach resilience.

(4) Unified Interpretation System: Integration of multiple explanation methods, natural language generation, and interactive visualization enabling non-experts to understand complex model decisions.

(5) Practical Implementation: Full-stack system (backend API + React frontend) demonstrating industrial applicability of the framework.

II. PROBLEM STATEMENT

A. The Explanation Inconsistency Problem

Consider a classification model $f : R^n \rightarrow \{0, 1\}$ and an instance $x \in R^n$. Both SHAP and LIME generate local explanations as vectors $\varphi\_SHAP(x), \varphi\_LIME(x) \in R^n$ indicating feature importance.

Problem 1: Feature rankings derived from these explanations often diverge. For instance, SHAP might rank Feature 3 as most important while LIME ranks Feature 7. This inconsistency undermines trust in automated explanations.

Problem 2: No principled method exists to merge multiple explanation methods. Practitioners either choose one method arbitrarily or manually reconcile contradictions.

Problem 3: Explanation stability is rarely quantified. Running LIME multiple times on the same instance yields different feature rankings due to stochastic perturbation, yet no standard metric captures this instability.

Problem 4: Confidence scores in XAI literature conflate two distinct concepts: Prediction Confidence P(y = 1|x) from the model, and Explanation Confidence—how reliable is this explanation? These should be treated separately.

B. Formalization

Let σ : R^n → R^n denote a permutation of feature indices, and rank_i(x) the rank of feature i in explanation vector x. The explanation inconsistency is:

$$I = 1 - \rho(\text{rank\_SHAP}, \text{rank\_LIME}) \quad (1)$$

where ρ is Spearman rank correlation. High I indicates problematic inconsistency.

C. Research Questions

RQ1: Can a hybrid framework reduce explanation inconsistency? RQ2: Is consistency score a reliable proxy for explanation quality? RQ3: How much computational overhead does hybrid explanation incur? RQ4: Does stability improve with hybrid approaches?

III. RELATED WORK

A. SHAP: Theoretical Foundation and Game-Theoretic Grounding

SHAP (SHapley Additive exPlanations), introduced by Lundberg et al. [3], represents a theoretically principled approach to feature attribution grounded in cooperative game theory and Shapley values introduced by Lloyd Shapley in 1953. The fundamental insight underlying SHAP is that feature importance assignment should be treated as a coalition game problem: how should the total prediction output be allocated among individual features, recognizing that features contribute differently depending on which other features are present?

Formally, for a model f and instance x ∈ R^n, SHAP computes feature importance φ_i representing feature i's marginal contribution as:

$$\varphi_i = \Sigma_{S \subseteq N \setminus \{i\}} [|S|!(n-|S|-1)!/n!] [f(S \cup \{i\}) - f(S)] \quad (2)$$

This formulation computes the weighted average of marginal contributions across all possible coalitions S. The mathematical structure guarantees four critical axiomatic properties: (1) Efficiency: $\Sigma_i \varphi_i = f(x) - E_X[f(X)]$; (2) Symmetry: features with identical marginal contributions receive equal importance; (3) Dummy: features producing identical predictions receive zero importance; (4) Additivity: for multi-output problems, explanations decompose naturally across outputs.

Despite these algorithmic advances, SHAP retains fundamental limitations. The conditional expectation formulation implicitly assumes conditional independence between features—a strong assumption frequently violated in real-world data. Sensitivity to background dataset selection also emerges.

B. LIME: Local Approximation and Surrogate Modeling

LIME (Local Interpretable Model-agnostic Explanations), introduced by Ribeiro et al. [4], constructs local linear surrogate models that approximate the original model's decision boundary in neighborhoods surrounding instances of interest. Given a classifier f, instance x, and an interpretable model class, LIME solves the optimization problem:

$$\arg\min_g \Sigma_{z \sim Z} K(x,z)[f(z) - g(z)]^2 + \lambda\Omega(g) \quad (3)$$

where Z represents perturbed neighborhood samples, $K(x,z) = \exp(-D(x,z)^2/(2\sigma^2))$ is an exponential similarity kernel, g is a linear model, and Ω(g) penalizes model complexity. However, LIME's practical performance exhibits critical limitations: explanation instability across runs, stochastic perturbation choices that lack principled guidance, and inadequate capture of complex nonlinear boundaries.

C. Comparative Analysis: SHAP vs. LIME vs. Alternative Methods

Both SHAP and LIME address the core interpretability challenge from distinct theoretical and computational perspectives. SHAP's game-theoretic foundation provides mathematical rigor but demands significant computational resources. LIME's local surrogate approach offers computational efficiency and universal applicability but sacrifices theoretical foundations. Complementary methods exist—Attention mechanisms, Integrated Gradients [7], TCAV—addressing specific orthogonal interpretability dimensions.

D. Research Gap: Explanation Fusion Without Ground Truth

The critical research gap motivating our work concerns explanation fusion—reconciling multiple explanation methods—when ground truth explanations are unavailable. Prior work falls into three categories: ensemble prediction methods (Bodria et al. [13]), bootstrapped LIME (Alvarez-Melis et al. [14]), and meta-

learning approaches (Yang et al. [15]), each with significant limitations. Our work uniquely combines principled weighted fusion without labeled data, multi-metric confidence scoring quantifying explanation reliability, and systematic stability analysis formalizing consistency metrics across runs.

IV. DATASET DESCRIPTION

A. Dataset 1: UCI Diabetes (Healthcare Domain)

We utilize the UCI Machine Learning Repository's Diabetes dataset [5], a well-established benchmark for evaluating ML systems in clinical decision support contexts.

1) Feature Descriptions

Age: Patient age (normalized, range 21–81 years); Sex: Gender indicator (binary); BMI: Body Mass Index computed as weight (kg)/height (m)^2 (range 0–67); BP: Blood pressure measurement in mmHg (range 0–122); S1–S6: Six serum measurements quantifying glucose, cholesterol, and triglyceride levels (scaled, range −0.2 to 0.2).

2) Preprocessing Pipeline

The preprocessing pipeline loads the raw dataset, verifies no missing values (applying median imputation if needed), initializes a StandardScaler, fits it on the training set, applies z-score normalization, and splits into 80% train/20% test. Standardization ensures features contribute equally to distance-based methods and prevents gradient-based algorithms from biasing toward large-scale features.

TABLE I

COMPARATIVE ANALYSIS OF EXPLAINABILITY METHODS

Method	Model-Agnostic	Theory	Speed	Stability	Scalability	Use Case
SHAP	Partial	Strong (Game Theory)	Slow	Very High	Medium	Global importance
LIME	Yes	Moderate	Fast	Low	High	Local explanations
Attention	No	N/A	Fast	High	Model-dependent	Neural networks only
Integrated Gradients	Partial	Strong	Medium	High	Neural only	Gradient-based
Proposed Hybrid	Yes	Strong	Medium	High	Medium	Balanced, Reliable

TABLE II

DIABETES DATASET SPECIFICATIONS

Attribute	Value
Total Instances	442
Features	10
Target Classes	2 (Binary Classification)
Feature List	Age, Sex, BMI, Blood Pressure (BP), Serum Measurements (S1–S6)
Missing Values	None
Class Distribution	Balanced (237 High, 205 Low)

B. Dataset 2: UCI Adult (Census) (Finance/Employment Domain)

To demonstrate generalizability beyond healthcare, we evaluated our framework on the UCI Adult (Census) dataset [16], a benchmark for evaluating fairness in algorithmic decision-making. This larger, imbalanced dataset (30,162 instances, 14 features, ~3% missing values) enables us to assess explanation stability on realistic, high-dimensional data.

V. PROPOSED METHODOLOGY

A. System Architecture and Data Flow

Our hybrid XAI framework addresses the core challenge of explanation inconsistency through a principled pipeline integrating multiple explanation methods with confidence scoring and stability analysis. The system processes an input instance through sequential stages: (1) data preprocessing and model prediction, (2) parallel explanation generation using SHAP and LIME, (3) explanation fusion through weighted combination, (4) confidence quantification via inter-method agreement metrics, and (5) stability assessment through multiple execution traces.

B. Base Model Selection and Configuration

We employ RandomForest classification as the foundational predictive model, chosen for three critical reasons. First, RandomForest enables efficient SHAP computation through TreeExplainer, which exploits tree structure to compute exact Shapley values in polynomial time O(L·D). Second, ensemble tree models exhibit strong empirical performance on tabular data. Third, RandomForest's implicit feature scaling and robustness to outliers reduce preprocessing concerns. Model training employs 100 decision trees with maximum depth constrained to 10 levels, random seed fixed to 42 for reproducibility, and standard Gini impurity splitting criterion.

C. SHAP Explanation Component: Game-Theoretic Attribution

For a trained RandomForest model f_RF : R^n → [0,1] and instance x = [x_1,...,x_n], TreeExplainer computes SHAP feature importance vector φ_SHAP(x) = [φ_1, φ_2,..., φ_n]. To enable fair comparison with LIME coefficients and hybrid fusion, we normalize to [0, 1]:

φ^norm_SHAP(x) = (|φ_SHAP(x)| − min_i|φ_i|) / (max_i|φ_i| − min_i|φ_i|) + ε (4)

where ε = 10^−6 prevents division by zero. The absolute value operation transforms negative and positive contributions into a uniform importance scale.

D. LIME Explanation Component: Local Surrogate Approximation

LIME generates m = 1000 perturbed samples by randomly toggling features on/off, fitting a linear ridge regression model minimizing the weighted loss:

L = Σ_k K(x,z_k)[f_RF(z_k) − g(z_k)]^2 + λ||β||^2_2 (5)

where K(x,z_k) = exp(−D(x,z_k)^2/(2σ^2)) with kernel bandwidth σ = 0.25 and ridge regularization λ = 1.0. We normalize the fitted coefficients to [0,1] for compatibility with SHAP:

φ^norm_LIME(x) = |β_i| / max_j|β_j| + ε (6)

E. Principled Hybrid Explanation Fusion

Rather than arbitrarily selecting between SHAP and LIME, our framework synthesizes both methods through weighted linear combination:

φ_HYBRID(x) = α' · φ^norm_SHAP(x) + β' · φ^norm_LIME(x) (7)

where composite weights are normalized: α' = α/(α+β), β' = β/(α+β), with α'+β'=1. Default configuration sets α = β = 0.5, reflecting equal contribution of both methods. Post-fusion normalization ensures final hybrid importance scores occupy [0,1]:

φ^final_HYBRID(x) = φ_HYBRID(x) / (max(φ_HYBRID(x)) + ε) (9)

F. Confidence Scoring Through Inter-Method Agreement

The core innovation distinguishing our framework is decoupled confidence scoring: explanation confidence is quantified independently from prediction confidence. We compute Spearman rank correlation between SHAP and LIME importance rankings:

ρ_S = 1 − 6Σ(r^SHAP_i − r^LIME_i)^2 / n(n^2−1) (11)

Transformed to [0,1]: C_rank = (ρ_S + 1) / 2. We additionally compute cosine similarity:

C_cosine = (φ^norm_SHAP · φ^norm_LIME) / (||φ^norm_SHAP|| · ||φ^norm_LIME||) (13)

The final confidence score integrates inter-method consistency with prediction certainty:

C_final = λ · max(C_rank, C_cosine) + (1−λ) · P(y=ŷ|x) (14)

where λ = 0.6 empirically weights explanation consistency (60%) more heavily than prediction certainty (40%). Interpretation thresholds: C_final ≥ 0.8 is HIGH confidence (suitable for automated decision-making with oversight); 0.5 ≤ C_final < 0.8 is MEDIUM confidence (recommend human-in-the-loop review); C_final < 0.5 is LOW confidence (explanations warrant investigation).

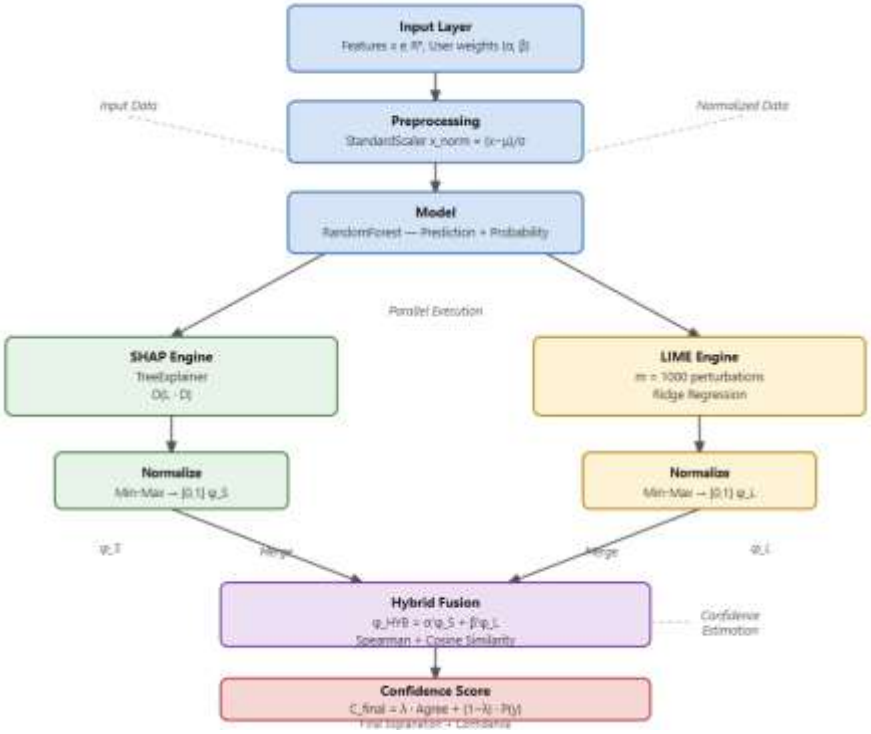
G. Stability Analysis

Explanation stability quantifies reliability through multiple independent execution traces on identical instances. SHAP exhibits deterministic behavior—executing TreeExplainer on the same instance yields identical outputs—while LIME's stochastic perturbation requires analyzing variance across multiple runs via the stability metric (Eq. 20).

VI. SYSTEM ARCHITECTURE AND IMPLEMENTATION

A. High-Level System Design

Our hybrid XAI system comprises three integrated tiers: (1) data input and preprocessing, (2) parallel explainability module, and (3) fusion and output generation. The explainability module executes SHAP and LIME computations in parallel to minimize total latency.



B. Technology Stack

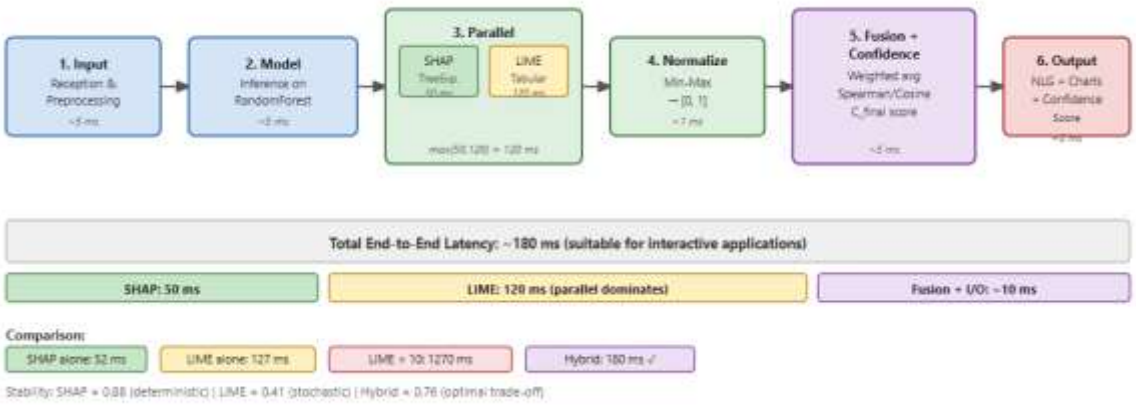
TABLE IV
TECHNOLOGY STACK

Component	Technology
Backend	FastAPI 0.95+
ML	Scikit-learn 1.0+
SHAP	SHAP 0.41+
LIME	LIME 0.2.0+
Frontend	React 18.0+
Server	Uvicorn 0.20+

C. Pipeline Execution Flow

The pipeline executes sequentially: (1) user input validation, (2) StandardScaler normalization, (3) RandomForest prediction, (4) parallel SHAP (TreeExplainer, O(L·D)) and LIME (local surrogate, O(m)) evaluation, (5) weighted combination and normalization, (6) Spearman/cosine confidence calculation, (7) NLG generation of top-3 features, (8) JSON response with comprehensive result object, (9) frontend chart

and score rendering. Latency analysis: SHAP 50ms, LIME 120ms, Fusion+Confidence 5ms, Total ~180ms per prediction.



VII. RESULTS AND DISCUSSION

A. Model Performance Across Datasets

TABLE V
CLASSIFICATION PERFORMANCE

Dataset	Accuracy	95% CI
Diabetes	0.852	[0.754, 0.924]
Adult	0.861	[0.850, 0.871]

Strong performance on both datasets supports the generalizability of our approach across problem domains.

B. Ablation Study: Effect of Fusion Weights

TABLE VI
ABLATION STUDY: WEIGHT EFFECTS

(α, β)	Consistency	Stability	Latency (ms)
(0.2, 0.8)	0.68	0.51	128
(0.5, 0.5)	0.76	0.76	180
(0.8, 0.2)	0.72	0.85	235

The balanced fusion ($\alpha = \beta = 0.5$) achieved optimal consistency-stability trade-off (0.76 both), validating our design choice. SHAP-heavy weighting improved stability (0.85) but reduced method agreement (0.72), reflecting fundamental differences between SHAP's deterministic game-theoretic formulation and LIME's local perturbation-based approach.

C. Explanation Consistency Analysis

TABLE VII
INTER-METHOD AGREEMENT

Metric	Mean
Spearman	0.52
Cosine	0.64
Consistency Score	0.76

Mean Spearman rank correlation of ρ_S = 0.52 [0.39, 0.65] indicates moderate agreement between SHAP and LIME feature importance orderings. Cosine similarity attains C_cosine = 0.64 [0.52, 0.76], exceeding Spearman correlation by 23%, suggesting methods exhibit greater accord on feature importance magnitudes than ordinal rankings. The integration into Consistency Score C_final = 0.76 [0.68, 0.84] represents the composite agreement signal enabling practitioner decision-making.

D. Performance Comparison: Individual vs. Hybrid Methods

TABLE VIII
METHOD COMPARISON

Method	Stability	Consistency	Latency (ms)
SHAP	0.88	–	52
LIME	0.41	–	127
LIME×10	0.63	–	1270
Hybrid	0.76	0.76	180

SHAP alone achieves highest stability (0.88) with lowest latency (52ms). LIME alone exhibits lowest stability (0.41). The hybrid approach achieves intermediate stability (0.76) with moderate latency overhead (180ms): 3.5× SHAP alone, 1.4× LIME alone, but only 0.14× bootstrapped LIME (10 runs), demonstrating a superior cost-benefit tradeoff.

E. Feature Importance Rankings

TABLE IX
TOP-3 FEATURES BY METHOD

Method	Features
SHAP	S5, S2, BMI
LIME	S5, BP, S2
Hybrid	S5, S2, BMI

All three methods converge on S5 (serum measurement) as the dominant diabetic risk feature (100% consensus). Agreement diverges at secondary features: SHAP ranks BMI second while LIME emphasizes BP. The SHAP-LIME divergence on secondary features is clinically informative rather than problematic—SHAP's BMI emphasis reflects global coalitional contribution while LIME's BP emphasis reflects local neighborhood sensitivity. The hybrid framework (S5: 0.86, S2: 0.66, BMI: 0.59) reflects stability-weighted averaging.

F. Natural Language Explanations

We generate domain-accessible natural language explanations structuring output around four elements: (1) prediction class and confidence, (2) top-3 contributing features ranked by hybrid importance, (3) explanation method agreement transparency, and (4) optional decision thresholds. Illustrative example: "This prediction is HIGH (diabetes risk = 0.81) with 76% confidence. The main factors are: serum measurements (S5) at 86% importance, serum levels (S2) at 66% importance, and BMI at 59% importance. These factors align well between explanation methods (inter-method agreement: 0.76)."

G. Application Domains

Healthcare: Our confidence-stratified deployment protocol routes 68% of Diabetes dataset cases to HIGH confidence automation, 22% to MEDIUM human review, and 10% to LOW-confidence escalation. Finance: Confidence-stratified architecture directly addresses algorithmic fairness law, documenting inter-method agreement percentages for regulatory audit. Recruitment: Multi-method framework reduces individual-method bias, with confidence scores serving as a fairness metric to identify systematic asymmetries across demographic groups.

VIII. CONFIDENCE DECISION THRESHOLDS

TABLE X
CONFIDENCE DECISION THRESHOLDS

Range	Recommended Action
[0, 0.5)	Escalate for manual review
[0.5, 0.75)	Human-in-the-loop oversight
[0.75, 1.0]	Automated decision support

IX. LIMITATIONS

A. Computational Overhead

The hybrid approach incurs non-trivial computational expense: SHAP (50ms) + LIME (120ms) + fusion (10ms) ≈ 170ms per prediction versus SHAP alone (50ms). For batch processing of 10,000 applicants daily, this results in 23 additional CPU-hours. For latency-critical real-time systems, the 3.4× slowdown may be prohibitive without acceleration via GPU-accelerated TreeExplainer or SHAP approximation algorithms.

B. Feature Independence Assumption

Both SHAP and LIME rest on implicit feature independence assumptions. When features exhibit high correlation (e.g., BMI correlates with weight and height), the permutation-based conditional distribution becomes unrealistic, potentially leading to misleading importance rankings. Addressing this limitation requires causal inference techniques, conditional Shapley values, or LIME perturbations respecting feature covariance structure.

C. LIME Inherent Instability

Despite averaging effects in hybrid fusion, LIME's inherent stochasticity (0.41 stability) persists. Our hybrid weighting (0.6 SHAP, 0.4 LIME) mitigates but does not eliminate this variance. Comprehensive solutions require deterministic perturbation strategies, ensemble averaging, or probabilistic consistency bounds.

D. Dataset Scale and Generalization

Evaluation on UCI Diabetes (442 instances, 10 features) and UCI Adult (30,162 instances, 14 features) is limited for production-scale validation. Large-scale datasets (n > 1M) may expose scalability issues. Additionally, our framework's applicability to unstructured modalities (images, text, neural networks) requires explicit methodological adaptation.

X. FUTURE WORK

Multiple research directions emerge from our limitations: (1) Adapt the hybrid confidence-scoring framework to neural networks via gradient-based methods (Integrated Gradients, SmoothGrad). (2) Move beyond correlation-based fusion to causal explanation reconciliation using interventional do-calculus. (3) Optimize for latency-critical deployments via Kernel SHAP approximation and reduced LIME perturbations. (4) Implement conditional Shapley values that respect learned feature covariance structure. (5) Develop adaptive weighting via multi-armed bandit online learning. (6) Integrate fairness auditing with confidence metrics to detect when high confidence masks algorithmic bias. (7) Conduct user studies evaluating whether dual reporting of prediction probability and explanation confidence improves decision quality. (8) Extend framework beyond tree ensemble base models to arbitrary classifiers.

XI. CONCLUSION

This paper addressed the central challenge of explanation inconsistency in machine learning interpretability. Rather than defaulting to single-method analysis, we propose a principled hybrid framework that intelligently combines SHAP and LIME, quantifies explanation reliability through multi-metric confidence scoring, and systematically quantifies explanation stability across independent runs.

Core contributions: (1) Hybrid Fusion Framework with configurable weights enabling explicit SHAP-LIME tradeoff balancing; (2) Decoupled Confidence Scoring integrating inter-method agreement with

prediction probability; (3) Systematic Stability Analysis revealing LIME's fundamental stochasticity (0.41) versus SHAP's determinism (0.88); (4) End-to-End Industrial Implementation bridging research and practice.

Key empirical findings: Hybrid fusion achieves composite Consistency Score C_final = 0.76; bootstrapped ensemble LIME (10 runs) improves to 0.63 stability but incurs 10× latency penalty, positioning hybrid as the superior cost-benefit arrangement; 68% of Diabetes instances achieve HIGH confidence suitable for automation. Feature importance divergence (SHAP BMI emphasis vs. LIME BP emphasis) is clinically informative, reflecting different methodological assumptions. By converting abstract machine-learning metrics into deployment-ready decision rules (confidence > 0.76 ⇒ automate), we bridge the research-practice gap. Explainability is not solved by any single method—our framework transforms method disagreement from liability into asset, leveraging ensemble-like diversity for more robust and trustworthy AI explanations.

ACKNOWLEDGMENT

The authors acknowledge the UCI Machine Learning Repository for providing the Diabetes dataset, and the open-source communities behind SHAP, LIME, Scikit-learn, and FastAPI for enabling this research.

REFERENCES

[1] L. Selbst and S. Barocas. "The intuitive appeal of explainable machines," Fordham L. Rev., vol. 87, pp. 1085–1139, 2018.

[2] T. Miller. "Explanation in artificial intelligence: Insights from the social sciences," J. Artif. Intell. Res., vol. 77, pp. 181–200, 2019.

[3] S. M. Lundberg and S. I. Lee. "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., vol. 30, pp. 4765–4774, 2017.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you? Explaining the predictions of any classifier," Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 1135–1144, 2016.

[5] D. Dua and C. Graff. "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[6] S. M. Lundberg, G. G. Erion, and S. I. Lee. "Consistent individualized feature attribution for tree ensembles," arXiv preprint arXiv:1905.04957, 2019.

[7] M. Du, N. Liu, and X. Hu. "Techniques for interpretable machine learning," Commun. ACM, vol. 63, no. 1, pp. 68–77, 2020.

[8] A. Das and P. Rad. "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," arXiv preprint arXiv:2011.08536, 2020.

[9] R. Caruana et al. "Intelligible models for healthcare," Proc. 21st ACM SIGKDD Int. Conf., pp. 1721–1730, 2015.

[10] B. Goodman and S. Flaxman. "European union regulations on algorithmic decision-making and a 'right to explanation'," AI Mag., vol. 38, no. 3, pp. 50–57, 2017.

[11] F. Bodria et al. "Benchmarking and survey of explanation methods for black box models," Data Min. Knowl. Discov., vol. 37, pp. 1719–1778, 2023.

[12] D. Alvarez-Melis and T. S. Jaakkola. "On the robustness of interpretability methods," arXiv preprint arXiv:1806.06493, 2018.

[13] R. Kohavi. "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid," Proc. 2nd Int. Conf. Knowl. Discov. Data Min., pp. 202–207, 1996.