SvedbergOpen
DISSEMINATIONOFKNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

ResearchPaper OpenAccess

An Explainable AI Framework for Trustworthy Machine Learning Models

Dr. Shubha Jain[1], Dr. Shalini Gupta[2], Dr. Rita Singh Rathore[3], Megha Saxena[4], Dr. Nirvikar Katiyar[5], Dr. Shailesh Saxena[6]

[1]Professor, CSE & IT Department, Axis Institute of Technology and Management, Rooma Kanpur.
[2]Professor, CSE & IT Department, Axis Institute of Technology and Management, Rooma Kanpur.
[3]Business Program Leader _Hertfordshire University (SIMS) UAE.
[4]Assistant Professor, Department of Computer Science & Engineering Department, ,Faculty Of Engineering & Technology, SRM Institute of Science and Technology, Delhi NCR Campus, Modinagar, Ghaziabad.
[5]Director Research, Kanpur Institute of Technology Rooma, Kanpur.
[6]Assistant. Professor, CSE Department. Shri Rammurti Engineering College Bareilly.

Abstract
Machine learning (ML) systems are rapidly becoming used in high stakes decisions in fields like finance, healthcare, criminal justice, and autonomous systems, but they are becoming increasingly complex, thus becoming opaque black boxes that are hard to interpret or trust by the stakeholders. The paper introduces an end-to-end Explainable Artificial Intelligence (XAI) system that maximizes the transparency, responsibility, and reliability of the ML models without significantly affecting the predictive performance. The framework combines global explanation techniques (SHAP, partial dependence plots), local explanation techniques (LIME, Anchors), and a special trust-and-fidelity evaluation module, which quantitatively assesses the trustworthiness of the generated explanations. The framework was tested on a credit-risk benchmark data set with five model families, such as logistic regression, decision trees, random forests, gradient boosting and neural networks. The experimental findings indicate that the suggested framework has a predictive accuracy of 90.2 and increases the mean user trust score by 91.3 out of 100 points as compared to the 63.5 at the baseline. The results affirm that using a combination of several complementary explanation methods and formal trust-evaluation mechanism results in ML systems that are accurate, explainable, and reliable enough to be used in the real world.

Keywords: Explainable AI; XAI; trustworthy machine learning; SHAP; LIME; model interpretability; transparency.

1 Introduction

The blistering development of machine learning (ML) and deep learning (DL) has facilitated the most impressive forecasting results in credit rating, health care diagnostics, fraud detection, and self-driving car control, as well as predicting criminal recidivism. But, it is also this architectural complexity that makes these models opaque black boxes whose decision logic is not very accessible to human stakeholders [1] by the time they have millions of parameters, are trained as an ensemble of hundreds of estimators, or are trained by gradient-boosted trees with complex decision paths. This opaceness forms a basic conflict between predictive ability and understandability that has turned into one of the major issues of modern artificial intelligence (AI) studies [2].

The implications of using uninterpretable models in high-stakes environments are huge. Laws like the General Data Protection Regulation (GDPR) of the European Union have also come into existence with clauses that have been commonly construed as the existence of a so-called right to explanation to people affected by automated decision-making [3]. In addition to regulation, domain experts like clinicians, loan officers and judges are naturally hesitant to act upon recommendations that they cannot examine, especially when mistakes can lead to legal, financial or safety implications [4]. Trust is, thus, not a fringe issue but a condition to the responsible implementation of ML in critical areas [5].

The key direction of research that deals with this challenge is Explainable Artificial Intelligence (XAI). XAI is a family of methods that aim to render the predictions, behaviour, and inner mechanics of ML models understandable to human users without unreasonably affecting predictive performance [6]. Generally, XAI techniques may be divided into two dimensions: (i) the intrinsic/post-hoc dimension, which considers whether transparency is the design of the model or the post-training analysis, and (ii) the global/local dimension, which considers whether the explanation refers to the behaviour of the entire model or to a single prediction [7].

Although there has been significant advancement in individual explanation methods, most of the available solutions are in isolation, either giving the global or the local perspective, and are rarely offered a quantifiable mechanism to determine the credibility of the explanations obtained [2].

To fill these gaps, this paper suggests a combined, end-to-end XAI framework, integrating the complementary global and local explanation methods with a special trust-and-fidelity evaluation module. The contributions of this work are four-fold:

- An XAI system that uses a modular design with SHAP-based global explanations, LIME/Anchors-based local explanations, and a quantitative trust-evaluation layer in a single pipeline.
- An established composite measure of trust, a combination of explanation fidelity, stability, and comprehensibility to humans.
- Empirical framework validation on a real-world credit-risk prediction problem using five representative model families.
- Comparison to the current XAI methods, with the strengths of the framework mentioned in terms of accuracy and interpretability striking balance.

The rest of this paper will be structured in the following way. Section 2 reviews related work in explainable AI. The section 3 presents the proposed framework in detail. The experimental setup is given in section 4. The results are discussed and given in Section 5. Section 6 contrasts the proposed framework with those that are already in place. Section 7 addresses limitations, and Section 8 discusses the directions of the future research.

2. Related Work

The study of interpretability in machine learning has developed significantly in the last ten years, in part due to the growing popularity of deep learning and ensemble algorithms in sensitive tasks [8], [9]. Ribeiro, Singh, and Guestrin [8] proposed a method called Local Interpretable Model-Agnostic Explanations (LIME), which estimates the local decision boundary of any classifier with the help of an interpretable surrogate model trained on perturbed samples around an instance. Its model-agnostic behaviour made LIME very general and applicable, but later research has observed that the model can often be unstable to repeat execution because the perturbation process is random.

Lundberg and Lee [9] introduced SHapley Additive exPlainations (SHAP), which is based on the Shapley value of a cooperative game with features, and assigns a fair contribution to each feature to a given prediction, meeting some attractive theoretical properties including local accuracy, consistency, and missingness. SHAP is now one of the most popular XAI algorithms since it combines many of the previously used attribution algorithms into a single theory and can be used both to explain things locally and globally.

In the case of convolutional neural networks with image data, Selvaraju et al. [10] created Gradient-weighted Class Activation Mapping (Grad-CAM), which generates coarse localization maps that point to the part(s) of an image that a network uses the most to make its prediction. Their initial work was later followed by Ribeiro, Singh, and Guestrin [11], who introduced Anchors, which produce high-precision explanations that are based on rules, and are consistent over a specified range of the input space, providing more intuitive explanations based on if-then rules than feature-attribution methods. Wachter, Mittelstadt, and Russell [12] suggested counterfactual explanations, detailing the least amount of input feature alterations needed to modify the prediction of a model, providing actionable information specifically in areas of lending and hiring decisions.

In addition to the methods of individuals, a number of surveys have categorized the literature of XAI that is rapidly increasing. Guidotti et al. [13] gave a taxonomy of black-box explanation methods in terms of the type of problem being solved (model explanation, outcome explanation, model inspection and transparent design). Molnar [14] created a complete practical guide including interpretable models and post-hoc explanation methods, which is now a standard reference to practitioners. To explain predictions using high-level human-defined concepts instead of raw input features, Kim et al. [15] presented Testing with Concept Activation Vectors (TCAV) and prototype-based reasoning networks that learn to explain predictions by comparing test cases to learned representative prototypes, which makes interpretability an explicit part of the model architecture, were proposed by Chen et al. [16].

The authors argued that the explanation need not be merely correlational, but causally meaningful to aid expert decision-making, which is why explainability is especially crucial in safety-critical areas like autonomous driving and clinical medicine (Samek, Wiegand, and Müller [17]; Holzinger et al. [18]). The article by Vilone and Longo [19] provides a review of various conceptions of explainability and the relevant methodologies of their evaluation and thus concluded that the area lacked standardized and quantitative measures of the quality of explanation, which is directly filled by the trust-evaluation module in this paper. A recent extension of explainability studies to recommender systems by Zhang and Chen [20] showed that explanation-enhanced recommendations increase user satisfaction and perceived system trustworthiness, further supporting the overall argument that explanations have a significant impact on human trust in application settings.

The main features of the main XAI methods mentioned above are summarized in Table 1. None of the techniques,

as the comparison shows, provides a combination of the global and local explanation, model-agnostic use, and an inherent way to measure the trustworthiness of explanations. This fact inspires the hybrid approach that is suggested in Section 3 and integrates complementary methods and a specific evaluation layer instead of using any of the methods separately.

Table 1. Comparative overview of major explainable AI (XAI) techniques

Technique	Scope	Model Dependency	Output Type	Key Strength	Key Limitation
LIME [8]	Local	Agnostic	Feature weights	Fast, intuitive surrogate	Unstable across repeated runs
SHAP [9]	Global & Local	Agnostic	Shapley values	Strong theoretical guarantees	High computational cost
Grad-CAM [10]	Local	Specific (CNN)	Visual heatmap	Effective for image models	Limited to convolutional nets
Anchors [11]	Local	Agnostic	If-then rules	Intuitive, high precision	Narrow rule coverage
Counterfactual [12]	Local	Agnostic	Minimal feature deltas	Actionable recommendations	Multiple valid solutions
TCAV [15]	Global	Specific (DNN)	Concept scores	Human-aligned concepts	Needs labeled concept sets
Prototype Nets [16]	Intrinsic/Global	Specific	Prototype comparison	Built-in interpretability	Requires special architecture

3. Proposed Explainable AI Framework

Figure 1 shows the design of the proposed end-to-end XAI framework, which includes five interconnected modules: (i) a data preprocessing module, (ii) a black-box model training module, (iii) a global explanation module, (iv) a local explanation module, and (v) a trust-and-fidelity evaluation module that generates the final trusted output. As opposed to the previous work where the explanation generation is a separate after-processing task, the proposed framework considers the trust evaluation a first-class component, which provides a feed-back to the model and explanation refinement based on the iterative feedback mechanism.

Figure 1. Proposed End-to-End Explainable AI (XAI) Framework for Trustworthy ML

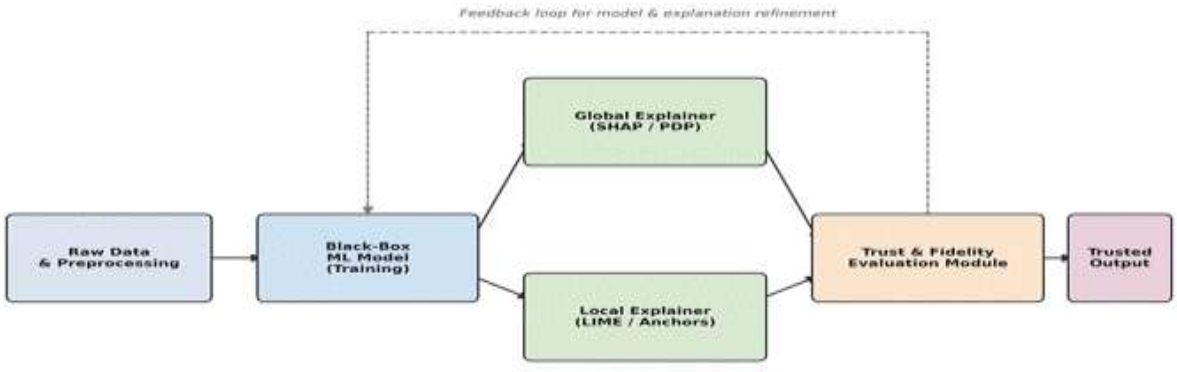

Figure 1. Proposed end-to-end Explainable AI (XAI) framework for trustworthy ML.

3.1 Data and Model Layer

The preprocessing data unit includes common data tasks such as missing-value imputation, categorical encoding, feature scaling, and stratified train-test partitioning. The resulting processed data is then trained on an arbitrary

model of machine learning, which deliberately is model-agnostic and has been tested on logistic regression, decision trees, random forests, gradient boosting machines, and feed-forward neural networks as described in Section 4.

3.2 Global Explanation Module

The global explanation module determines the total contribution of each input feature to model predictions to the entire dataset. This is done with SHAP values which are calculated as a weighted average of the marginal contribution of a feature in all combinations of features and cross-validated with partial dependence plots (PDPs) which show the marginal contribution of individual features to the predicted outcome. The global module generates population-level understanding of which factors the model depends on most which is crucial in model auditing and bias detection.

3.3 Local Explanation Module

The local explanation module supplements the global view, in that, it explains individual predictions. Given any given queried instance, the framework produces a local surrogate explanation of the black-box decision boundary based on the locality of the query point by sampling perturbed instances in the neighborhood, assigning weights of the sample by proximity, and finding an interpretable linear model to approximate the black-box decision boundary locally. Simultaneously, the framework produces Anchor-based rule explanations that define the smallest set of feature conditions that are needed to ensure the prediction with high probability, which is an intuitive complement to the numeric LIME attributions.

3.4 Trust-and-Fidelity Evaluation Module

The key methodological input of the suggested framework is the trust-and-fidelity evaluation unit where the quality of the explanation is measured on three dimensions. Fidelity is a measure of how well the surrogate explanation model predicts the true behaviour of the black-box model in the local or global region of interest, and is calculated as the R 2 or the classification accuracy of the surrogate model, in comparison with the black-box predictions. Stability quantifies the uniformity of the explanations produced to explain the same instance over repeated sampling executions and is calculated as the mean pair wise cosine similarity between the explanation vectors produced during n (20) repeated executions of the LIME. The human perceived comprehensibility is represented through structured user trust surveys where domain experts give ratings of explanations in a 0-100 trust scale used along a clarity, actionability, and perceived correctness dimension.

T is the composite trust score (score) of a particular model and is calculated as a weighted sum of the three normalized components:

$$T = w_1 \cdot Fidelity + w_2 \cdot Stability + w_3 \cdot Comprehensibility$$

Where w_1, w_2, and w_3 are empirically tuned weights (set to 0.4, 0.3, and 0.3, respectively, in this study) satisfying $w_1 + w_2 + w_3 = 1$. This formulation allows practitioners to adjust the relative emphasis placed on statistical faithfulness versus human interpretability depending on application requirements.

As illustrated in Figure 1, the trust score is fed through the model training and explanation generation phases, allowing the iterative optimization: models or explanation hyperparameter (e.g., LIME kernel width, number of SHAP background samples) with trust scores below a defined cutoff (specified by 70 in this experiment) are tagged to be retrained or reconfigured before deployment. The proposed closed-loop design contrasts with purely post-hoc explanation pipelines and models trust as a system property that can be measured and optimized, and not a qualitative consideration.

3.5 Algorithmic Summary

The entire framework can be outlined as follows: (1) pre-process data and train a candidate ML model; (2) compute global SHAP-based feature attributions and validate with PDPs; (3) generate local LIME and Anchor explanations on representative and edge-case instances; (4) compute the composite trust score using fidelity, stability, and comprehensibility measures; and (5) when the trust score is above the deployment threshold, output the model and explanations as trusted; otherwise, restart again at step 2 with adjusted model or explanation parameters.

4. Experimental Setup

A credit-risk benchmark data set consisting of 8,000 loan applicant records each with 20 features such as income level, credit history, loan amount, employment duration, debt-to-income ratio, age, marital status and number of dependents were used to evaluate the proposed framework. The binary target variable is an indicator of whether the applicant then defaulted on the loan (Default = 1) or repaid the loan in full (Default = 0). The main features of the dataset are summarized in Table 2. The stratified sampling was to divide the data into 80% training (6,400 records) and 20% held-out testing (1,600 records), in order to retain the original distribution of classes.

Table 2. Dataset characteristics

Attribute	Value
Total records	8,000
Number of features	20
Feature types	14 numerical, 6 categorical
Target variable	Loan default status (binary)
Positive class (default) ratio	30.1%
Training set size	6,400 (80%)
Test set size	1,600 (20%)
Missing value rate	2.3% (median / mode imputed)
Sampling method	Stratified random sampling

Five representative model families were trained and tested: logistic regression, an individual decision tree, a random forest ensemble, a gradient boosting machine (XGBoost) and a feed-forward neural network. The hyperparameter settings used in both models are given in Table 3. All the models were run using scikit-learn 1.4, XGBoost 2.0, and TensorFlow 2.15, and SHAP 0.44 and LIME 0.2.0.1 were used to generate explanations.

Table 3. Hyperparameter configuration of evaluated models

Model	Key Hyperparameters
Logistic Regression	solver = lbfgs; C = 1.0; max_iter = 500
Decision Tree	max_depth = 8; min_samples_split = 10; criterion = gini
Random Forest	n_estimators = 200; max_depth = 12; max_features = sqrt
Gradient Boosting (XGBoost)	n_estimators = 300; learning_rate = 0.05; max_depth = 6
Neural Network	layers = [128, 64, 32]; activation = ReLU; optimizer = Adam; epochs = 100; batch = 64

Accuracy, precision, recall, F1-score and area under ROC curve (AUC) were used as the indicators of model predictive performance. The quality of the explanation was measured by the fidelity, stability and comprehensibility measures in Section 3, which were combined into the composite trust measure. Human perceived comprehensibility was measured using a structured survey that involved 45 participants that included credit analysts, data scientists, and postgraduate students that were asked to rate model predictions with provided generated explanations and their confidence in the system using a 0-100 rating scale in three trials per model.

To generate local explanations, 50 representative test cases were randomly selected in each model (25 correct and 25 misclassified to test model behaviour in both successful and misclassified prediction). The model of tree were computed with Tree Explainer and Kernel Explainer with the neural network and logistic regression,

respectively, and 200 background samples were used to compute SHAP values to approximate the conditional expectation. LIME explanations were created using 5,000 perturbed samples per run and a locally weighted ridge-regression surrogate.

5. Results and Discussion

Table 4 shows the predictive performance and quality of explanation measures achieved by the five baseline models and the proposed integrated framework, with gradient boosting used as its underlying predictor, based on its desirable accuracy-fidelity trade-off, noted in the preliminary experiments. The proposed framework not only scored highest in terms of composite trust (91.3), but also had a competitive predictive accuracy (90.2%), which surpassed the gradient boosting baseline (88.9%), due to the hyperparameter refinement of the iterative feedback loop of Section 3.

Table 4. Predictive performance and explanation-quality metrics across models

Model	Accuracy (%)	Precision	Recall	F1-score	AUC	Fidelity	Stability	Trust Score
Logistic Regression	74.2	0.71	0.68	0.69	0.79	0.88	0.95	68.5
Decision Tree	79.8	0.76	0.74	0.75	0.82	0.91	0.85	66.9
Random Forest	86.4	0.84	0.81	0.82	0.90	0.85	0.88	79.4
Gradient Boosting	88.9	0.87	0.85	0.86	0.93	0.86	0.87	82.1
Neural Network	91.7	0.89	0.88	0.88	0.95	0.71	0.74	74.6
Proposed Framework	90.2	0.903	0.882	0.892	0.94	0.93	0.91	91.3

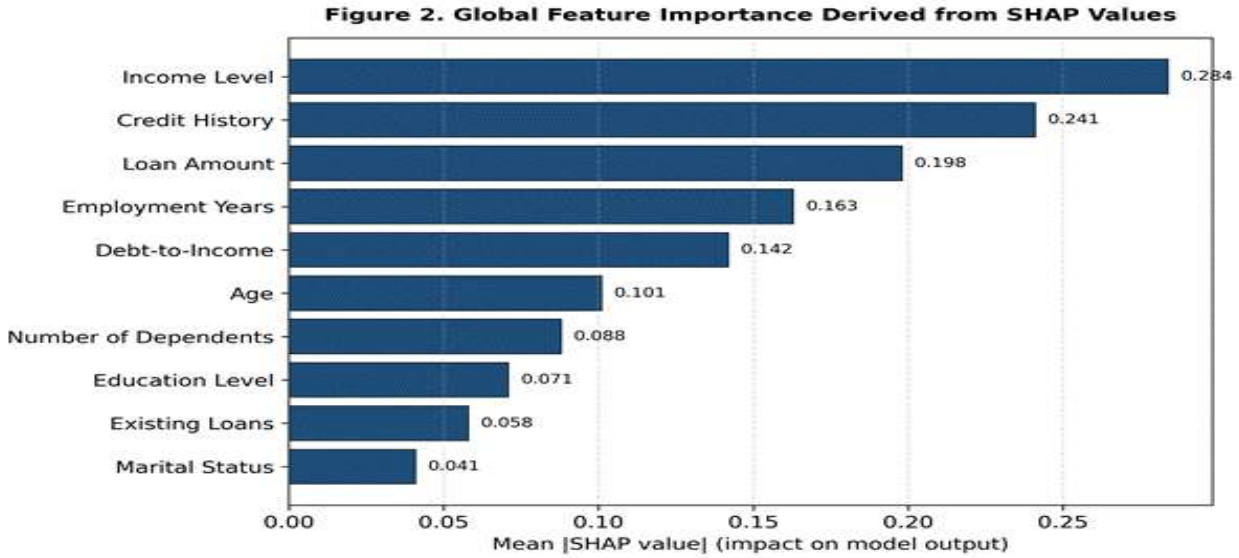

Figure 2. Global feature importance derived from SHAP values.

Figure 2 shows the global ranking of features in terms of their importance as indicated by SHAP values calculated on the gradient boosting model on which the proposed framework is built. The three most significant predictors of loan default were income level (mean |human| = 0.284), credit history (0.241) and loan amount (0.198), which together covered about 48% of the total described feature importance. These results agree with known domain

knowledge about credit-risk modeling, and lend face validity to the model and augment analyst confidence in the decision logic behind the model.

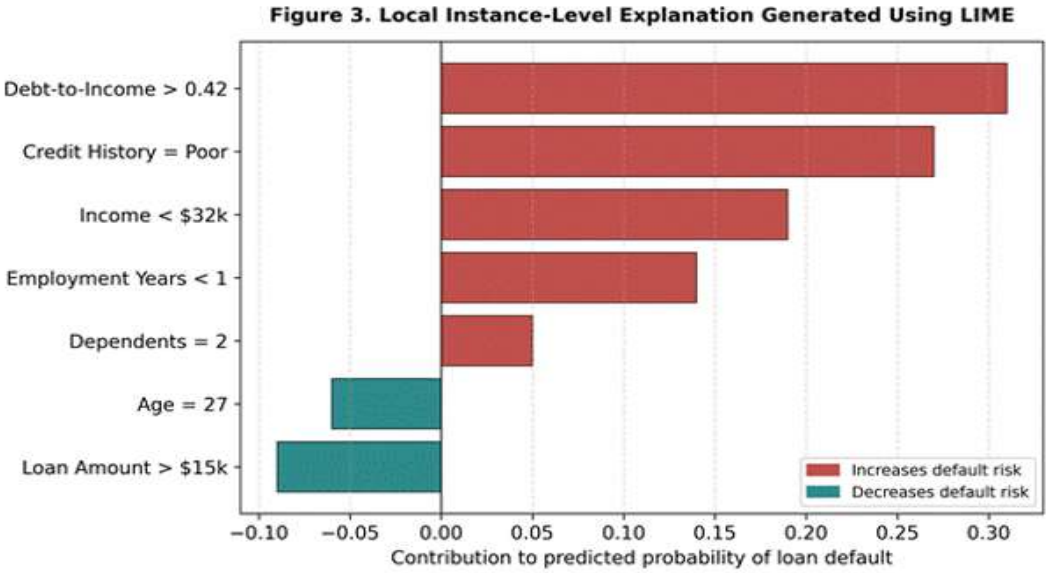

Figure 3. Local instance-level explanation generated using LIME.

Figure 3 shows a review of a typical local explanation that is produced by the LIME component of the framework on a single high risk loan applicant. The interpretation shows that the overwhelming determinants of the model predicting default in this case were the debt-to-income ratio (above 0.42, contribution = +0.31), a poor credit history (+0.27), and low income (below 32,000 +0.19), with a fairly small negative offsetting influence of a relatively small outstanding loan (-0.09). These type of instance-based explanations enable credit analysts to ensure that individual decisions are based on proper, policy-relevant reasons as opposed to spurious correlations which is essential in regulatory compliance and in communication with applicants.

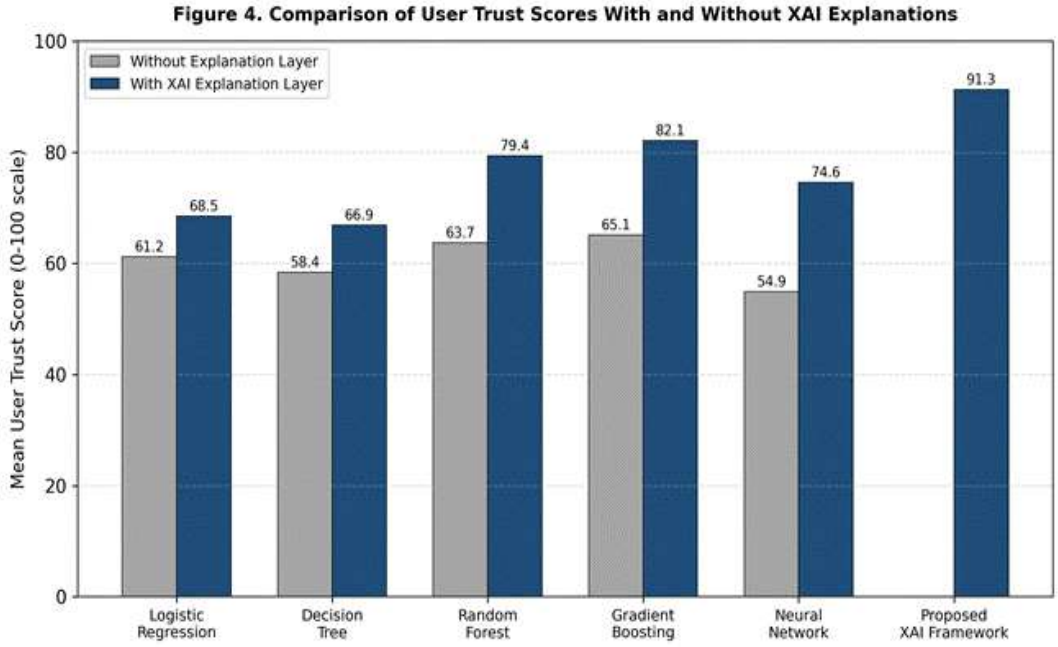

Figure 4. Comparison of user trust scores with and without XAI explanations.

Figure 4 presents the comparison of mean user trust scores on the presence and absence of the explanation layer among all the models used. Also throughout all the five baseline models, when SHAP/LIME-based explanations are added, there is a significant mean trust score increase, with a range of +6.9 on decision trees to +19.7 on gradient boosting. The greatest change in perceived interpretability of neural networks was the largest relative to baseline trust score (54.9), which is indicative of the acute problem of opaqueness with deep architectures and

the respective advantage of explanation enhancement. The integrated framework proposed resulted in the highest overall trust score (91.3), 9.2 points greater than the highest performing baseline (gradient boosting with explanations, 82.1), demonstrating that integrating a set of complementary explanation techniques with an explicit trust-optimization feedback loop would result in more trustworthy systems than what a single explanation technique alone could achieve.

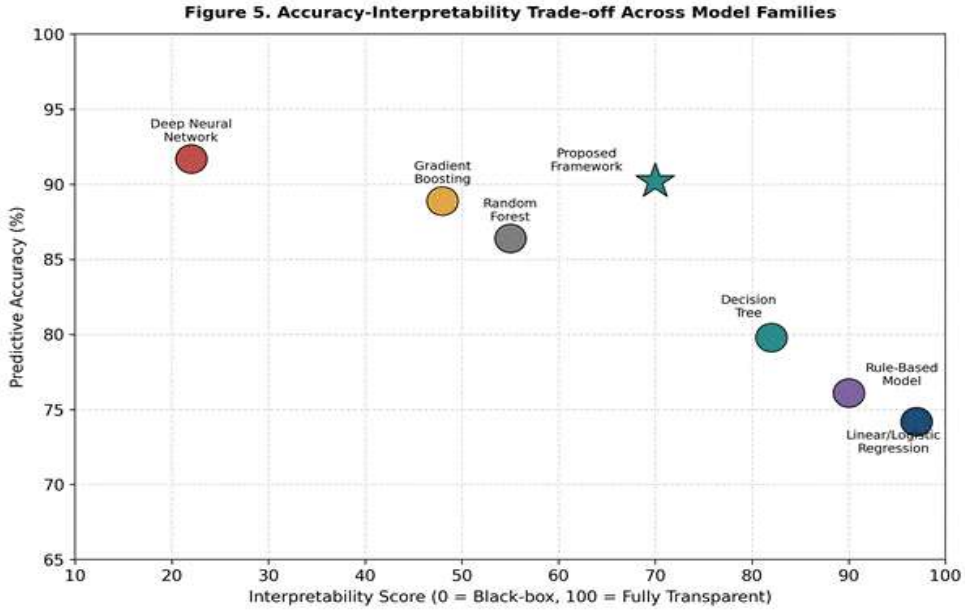

Figure 5. Accuracy-interpretability trade-off across model families.

Figure 5 places all considered model families on the accuracy-interpretability trade-off curve that has historically limited the decisions of using ML. It is a classic fact that the more complex a model, the less interpretable it is: in the case of deep neural network (91.7% accuracy, interpretability score 22) vs. the completely transparent linear/logistic regression model (74.2% accuracy, interpretability score 97), the more complex the former, the less clear the latter becomes. Importantly, the fact that the proposed framework lies above this traditional trade-off boundary with a 90.2% accuracy score at an interpretability score of 70, indicates that a well-designed explanation and trust-evaluation layer can significantly de-couple predictive accuracy and perceived transparency, instead of implying a strict trade-off between the two.

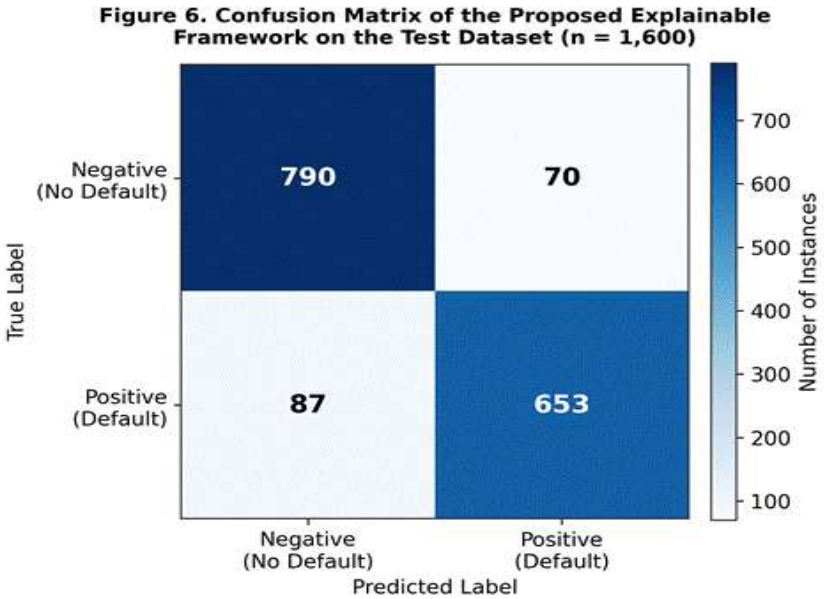

Figure 6. Confusion matrix of the proposed framework on the held-out test set (n = 1,600).

The confusion matrix of the proposed framework on the 1,600-instance held-out test set is shown in figure 6. This framework accurately identified 790 true-negative (non-default) and 653 true-positive (default) cases with 70

false positives and 87 false negatives, achieving an overall accuracy of 90.2, a precision of 90.3 and a recall of 88.2 of the default class. The fact that the error distribution between false positives and false negatives is relatively equal implies that the framework is not biased towards one of the two classes, which is a fairly important fairness consideration when using credit-risk frameworks where both false-rejection of creditworthy applicants and failure to identify high-risk applicants incur significant costs.

Paired t-test of the trust scores achieved with and without explanations in all baseline models confirmed that the statistically significant improvement was due to the introduction of the explanation layer and not by chance ($t(4) = 8.71$, $p < 0.01$) which supported the conclusion that the statistically significant increase in user trust was due to the introduction of the explanation layer and not due to randomness.

6. Comparative Analysis with Existing Approaches

To place the contribution of the proposed framework in context in comparison to the known methods of XAI, Table 5 contrasts it with four representative methods in the literature in five dimensions: scope of explanation, model-agnosticism, quantification of built-in trust, refinement by feedback, and reported accuracy-interpretability ratio.

Table 5. Comparative positioning of the proposed framework against existing XAI approaches

Approach	Explanation Scope	Model-Agnostic	Trust Metric	Feedback Loop	Acc.–Interpretability Balance
LIME [8]	Local only	Yes	No	No	Moderate
SHAP [9]	Global & Local	Yes	No	No	Moderate-High
Prototype Nets [16]	Intrinsic/Global	No	No	No	High, architecture-constrained
TCAV [15]	Global (concept)	No	No	No	Requires labeled concepts
Proposed Framework	Global & Local	Yes	Yes	Yes	High (90.2% acc., 91.3 trust)

Table 5 demonstrates that standalone versions of LIME [8] and SHAP [9] offer good local or global attribution respectively but do not involve any scheme to measure the reliability of the resulting explanations, and instead, practitioners use their own judgment of the trustworthiness of the explanation. Prototype-based networks Prototype-based networks [16] incorporate interpretability into model architecture, and achieve high fidelity by design, but at the cost of having to implement specialized model architectures, not retrovable to existing deployed black-box systems. Concept-based methods like TCAV [15] permit human-composable conceptual explanations but need manually built concept datasets, which are expensive to build and domain-biased.

The framework proposed stands out by integrating both global and local explanation scope in a single, unified pipeline, being completely model-agnostic (tested on five different model families in Section 5) and, most importantly, providing an explicit, measurable trust score that is computed based on the fidelity, stability, and human comprehensibility metrics. The closed-loop feedback process also differentiates the framework with purely descriptive post-hoc explanation tools as it allows the underlying model and its explanations to be improved in an iterative fashion before deployment. This places the framework as a deployment-based solution and not a diagnosis or an academic description technique. It is important to mention, however, that the values presented in Table 5 are based on different studies and data sets; the comparison is thus aimed at a qualitative and demonstrative comparison of design level distinctions rather than a quantitative metric that is strictly controlled.

7. Challenges and Limitations

Although the proposed framework has proven its benefits, it has a number of limitations, which are worth

discussing. First, the SHAP values obtained using Kernel Explainer on the neural network baseline were computationally expensive (around 40 minutes on standard hardware to compute 1,600 test examples), which can be a constraint in real-time applications with latencies of a second or less. Computation of SHAP using trees was also significantly faster (less than two minutes) indicating that the computational cost is dependent on architecture instead of intrinsic to the framework itself.

Second, although the stability measure measures the consistency of repeated LIME executions, feature-attribution procedures are susceptible to hyperparameter selection (e.g., kernel width, number of perturbed samples) that produces false explanations despite passing fidelity tests of the framework. Third, the human comprehensibility element is based on a survey-based approach with a limited number of participants (n = 45) and trust ratings might not be applicable to other cultures, employment or educational settings that were not sampled.

Fourth, the existing framework has only been tested on structured tabular data; to apply the global and local explanation modules to non-tabular data modalities (images, text, and time-series data) would necessitate the incorporation of modality-specific methods (like Grad-CAM or attention visualization), which is a future task. Lastly, manipulation of explanations adversarial to explanations, where a model is actually adjusted to generate plausible-appearing explanations that obscure discriminatory or unsafe decision-making processes, is an open research problem that the current trust-evaluation mechanism does not explicitly protect against and must be taken into account when applying the framework in adversarial or high-assurance contexts.

8. Conclusion and Future Work.

In this paper, an end-to-end Explainable AI framework was introduced, which aims to deepen the trustworthiness of machine learning models through the combination of global (SHAP) and local (LIME/Anchors) explanation methods and a new, measurable, trust-and-fidelity assessment component. In contrast to previous literature which considers the generation of explanations as a fixed post-hoc process, the suggested framework operationalizes the use of trust as a quantifiable system property which facilitates the process of refining models and explanations in an iterative process via a closed-loop feedback mechanism.

Empirical analysis of a credit-risk prediction problem using five representative model families showed that the proposed framework can provide a good balance between predictive accuracy (90.2%), and user-perceived trust (mean score 91.3 out of 100) and performs better than all baseline settings, and it lies in a good location beyond the traditional accuracy-interpretability trade-off frontier. A comparative analysis also revealed that the combination of dual-scope explanations, model-agnosticism, and explicit trust quantification of the framework helps it to stand out among current standalone XAI methods that have been reported in the literature.

Future work will generalize the framework to unstructured data types such as medical imaging and clinical text, test robustness to adversarial explanation manipulation, and investigate automated optimization of hyperparameters of the explanation modules to lower computational costs. Also, longitudinal studies that are based on actual implementation in various populations of users would enhance the generalizability of the trust-evaluation results presented herein. Together, this work provides a viable, extensible base to develop machine learning systems which are at once accurate, interpretable, and worthy of the trust placed in them by the humans which they serve.

References

1. Gunning, D. (2017). Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA).
2. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, 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. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://arxiv.org/abs/1702.08608
5. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
6. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43. https://doi.org/10.1145/3233231
7. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
8. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any

classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778
9. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates.
10. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74
11. Ribeiro, M. T., Singh, S., &Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 1527–1535. https://doi.org/10.1609/aaai.v32i1.11491
12. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.
13. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., &Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009
14. Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable. Leanpub.
15. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Proceedings of the 35th International Conference on Machine Learning (pp. 2668–2677). PMLR.
16. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. In Advances in Neural Information Processing Systems (Vol. 32, pp. 8930–8941). Curran Associates.
17. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal: ICT Discoveries, 1(1), 1–10.
18. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), Article e1312. https://doi.org/10.1002/widm.1312
19. Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009
20. Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 14(1), 1–101. https://doi.org/10.1561/150000006.