International Journal Of Management And Economics Fundamental

VOLUME Vol.05 Issue 10 2025
PAGE NO. 63-68

Bridging the Cognitive Gap: A Comparative Analysis of Contrastive and Feature-Based Explainability in High-Stakes Artificial Intelligence Systems

Dr. Melorina V. Strakhovskaya

Independent Researcher, Behavioral Economics & Explainable Financial AI, Saint Petersburg, Russia

Dr. Kostelia P. Vorontsenko

Independent Researcher, FinTech Risk Assessment & Interpretable Credit Models, Kazan, Russia

Received: 02 October 2025; **Accepted:** 16 October 2025; **Published:** 31 October 2025

Abstract: Background: As Artificial Intelligence (AI) systems, particularly Deep Neural Networks (DNNs), achieve superhuman performance in medical diagnostics and financial risk assessment, their inherent opacity—the "Black Box" problem—remains a critical barrier to adoption. Stakeholders in high-stakes domains require not just accurate predictions, but intelligible justifications that align with human cognitive reasoning.

Methods: This study provides a comparative evaluation of prominent Explainable AI (XAI) frameworks, specifically focusing on the dichotomy between feature-attribution methods (LIME, SHAP) and contrastive explanation approaches. We analyze these methodologies against a framework of "explanation effectiveness," assessing criteria such as local fidelity, consistency, and cognitive alignment with human decision-makers.

Results: Our analysis suggests that while feature-additive models like SHAP provide mathematical consistency in attributing contribution scores to input variables, they often fail to provide the causal intuition required in clinical settings. Conversely, contrastive explanations, which highlight "pertinent negatives" (what is missing but should be present for a different outcome), demonstrate higher efficacy in facilitating user trust and actionable insight, despite higher computational costs.

Conclusion: The transition from Black Box to "Glass Box" models is not merely a technical challenge but a socio-technical one. We conclude that for XAI to succeed in high-stakes environments, future architectures must prioritize contrastive reasoning that mirrors the differential diagnosis process used by human experts, moving beyond simple feature highlighting toward semantic intelligibility.

Keywords: Explainable AI (XAI), Medical Artificial Intelligence, Black Box Models, SHAP, LIME, Contrastive Explanations, Trustworthy AI

1. INTRODUCTION

The rapid proliferation of Artificial Intelligence (AI) across diverse sectors has fundamentally altered the landscape of decision-making. In domains characterized by high stakes, such as healthcare and financial services, the ability of an algorithm to predict an outcome is no longer sufficient; the algorithm must also be able to justify its reasoning. This requirement has given rise to the field of Explainable AI (XAI), which seeks to bridge the chasm between the opacity of advanced machine learning models—often termed "Black Boxes"—and the human need for transparency and interpretability, or the "Glass Box" ideal [4].

The urgency of this transition is driven by the complexity-accuracy trade-off. Historically, simpler models like linear regression or decision trees offered high interpretability but limited predictive power for complex, non-linear problems. Conversely, modern Deep Learning architectures, such as Convolutional Neural Networks (CNNs) used in medical imaging, offer superior accuracy but function as opaque mathematical transformations, encompassing millions of parameters that are unintelligible to human operators [5]. In a clinical setting, a physician cannot ethically rely on a diagnosis provided by a system that

cannot explain the pathological features driving its conclusion. As noted by Sheu and Pardeshi, the integration of AI into medical workflows is contingent upon the system's ability to provide explanations that foster trust and facilitate the verification of results [1].

Furthermore, the pressure for explainability is not merely functional but regulatory. Frameworks such as the European Union's General Data Protection Regulation (GDPR) have introduced the concept of a "right to explanation," mandating that individuals subject to automated decisions have the right to understand the logic involved. This is particularly pertinent in the financial sector, where credit scoring models must be non-discriminatory and transparent [12].

However, the definition of what constitutes a "good" explanation remains elusive. Is it a mathematical proof of feature importance? Is it a counterfactual scenario? Tjoa and Guan emphasize that medical XAI must move beyond technical metrics and address the semantic gap between computational output and clinical reasoning [2]. This study aims to dissect this challenge by comparing the efficacy of feature-based explanation methods, such as Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP), against contrastive frameworks. By analyzing these approaches through the lens of human cognitive requirements, we seek to determine which methodologies best support the transition from opaque prediction to transparent collaboration between human and machine.

2. Methodology

To rigorously evaluate the efficacy of current XAI methodologies, this study adopts a multi-dimensional evaluation framework that prioritizes "Explanation Effectiveness" over raw computational performance. This approach is grounded in the taxonomy proposed by Jung et al., which suggests that the essential properties of XAI in healthcare must include consistency, fidelity, and human-readability [3].

2.1 Evaluation Framework

We categorize our analysis into three primary dimensions:

1. Local Fidelity: The extent to which the explanation accurately reflects the model's behavior in the immediate vicinity of the instance being predicted. This is crucial for non-linear models where global interpretability is often impossible.

2. Cognitive Alignment: The degree to which the explanation format matches human reasoning processes. This involves assessing whether the explanation provides causal insight or merely correlational data.

3. Computational Feasibility: The latency and resource cost associated with generating post-hoc explanations, which is a critical factor for real-time applications like emergency medicine or high-frequency trading.

2.2 Algorithm Selection and Theoretical Context

The study focuses on three dominant classes of XAI techniques. First, we examine LIME, which approximates black-box models with simple, interpretable local surrogates (like linear models) to explain individual predictions [6]. Second, we analyze SHAP, a game-theoretic approach that assigns each feature an importance value for a particular prediction [12]. Finally, we investigate Contrastive Explanation Methods (CEM), which utilize pertinent positives and negatives to describe what minimal changes would be required to alter a decision [8].

The comparative analysis is contextualized within two high-stakes environments:

- Medical Diagnostics: Utilizing the context of a Recurrent Neural Network (RNN) or CNN predicting disease progression, as discussed by Ferreira et al. [10]. Here, the goal is to explain why a specific patient was flagged for high risk of mortality or disease onset.

- Financial Risk Assessment: Utilizing the context of credit scoring, where the rejection of a loan application must be justified to the applicant and regulators.

2.3 Analytical Procedure

Our methodology involves a theoretical synthesis of performance metrics reported in recent literature, combined with a qualitative assessment of the "Trust" factor. We rely on the distinction made by Ribeiro et al. between "trusting a prediction" (believing the specific output) and "trusting a model" (believing the system is generally sound) [6]. We evaluate how each XAI method contributes to these two distinct layers of trust. By simulating the interaction between these algorithms and a hypothetical domain expert (e.g., a radiologist or loan officer), we can identify the functional limitations of current "Glass Box" attempts.

3.Results

The analysis reveals significant divergences in how different XAI methods achieve transparency, highlighting a critical trade-off between mathematical axiom and human intuition.

3.1 Feature-Additive Methods: SHAP and LIME

Our review of LIME indicates that while it is highly effective at identifying the specific features in an image or dataset that contributed to a prediction, it suffers from stability issues. Because LIME relies on random sampling to create a local surrogate model, different runs of the algorithm on the same data point can occasionally yield slightly different explanations. In a medical context, such instability can be fatal to trust; a doctor cannot rely on a system that fluctuates in its justification [11].

SHAP addresses the consistency problem by leveraging Shapley values from cooperative game theory. It provides a robust, globally consistent method for feature attribution. In the context of credit risk, as explored by Gramegna and Giudici, SHAP values effectively highlight which financial history variables pushed a credit score up or down [12]. However, the result is often a dense vector of contributions—e.g., "Age contributed +0.02, Income contributed +0.15." While mathematically sound, this form of explanation imposes a high cognitive load on the user. It tells the user what happened, but not necessarily why it happened in a causal sense, nor does it intuitively suggest how to change the outcome.

3.2 The Cognitive Superiority of Contrastive Explanations

This leads to the most significant finding of our analysis: the misalignment between feature-additive explanations and human cognitive preference for contrastive reasoning.

3.2.1 The Psychology of "Why P rather than Q"

Human beings rarely ask "Why?" in an absolute sense. Instead, they ask "Why outcome P instead of outcome Q?" This is the foundation of contrastive explanation. When a physician diagnoses a patient, they are implicitly ruling out other conditions. When a bank denies a loan, the applicant wants to know what they need to change to get approved, not just why they were denied.

Feature-based methods (like SHAP) provide a "complete" account of the evidence, but this often leads to information overload. Dhurandhar et al. formalized this by distinguishing between Pertinent Positives (PP) and Pertinent Negatives (PN) [8].

- Pertinent Positives are the factors present in the input that support the current classification. For example, in an MRI scan identifying a tumor, the PP would be the pixels corresponding to the mass.

- Pertinent Negatives are the factors that are missing from the input but whose absence is necessary for the classification. Alternatively, they can be viewed as the minimal features that, if added, would change the classification.

Our analysis suggests that Pertinent Negatives are often more valuable for "Glass Box" transparency in high-stakes fields. For instance, in a medical diagnosis of a healthy patient, a feature-based explanation might confusingly highlight "normal" organs as contributing to the "healthy" label. A contrastive explanation, however, would simply state: " The patient is classified as healthy because there is no evidence of consolidation in the lower lobes." This aligns with the "explanation based on the missing" paradigm [8], which mirrors the "ruling out" process in clinical differential diagnosis.

3.2.2 Counterfactuals and Actionability

Extending the contrastive approach, counterfactual explanations provide actionable insights. In the financial domain, telling a user "Your loan was denied because your income is $40k" is less helpful than "Your loan would have been approved if your income was $45k." This shift from attribution to counterfactual generation transforms the AI from a passive judge into an active advisor.

However, generating these explanations is computationally intensive. Finding the minimal change required to flip a prediction in a high-dimensional space (like a 3D medical image) requires complex optimization that can introduce latency. Despite this, the gain in "explanation effectiveness" appears to outweigh the computational cost for high-stakes decisions where accuracy and trust are paramount.

3.2.3 The "Glass Box" Spectrum in Clinical Workflows

The transition from Black Box to Glass Box is not binary. Our analysis of recurrent neural networks in predictive medicine [10] indicates that "interpretable" architectures (like attention mechanisms) can offer a

middle ground. Attention maps allow clinicians to see which time-steps in a patient's history the model focused on. While this is a form of "Glass Box" design, it is often insufficient on its own. Attention implies correlation, not causation. Therefore, we observe that the most effective systems utilize a hybrid approach: using an inherently interpretable architecture (like attention-based RNNs) supplemented by post-hoc contrastive explanations to verify the reasoning.

3.3 Comparative Analysis of Post-Hoc Interpretability Methods in Complex Domains

To fully understand the implications of these findings, we must expand our view beyond the basic mechanics of these algorithms and analyze their deployment in complex, real-world ecosystems. The dichotomy between "feature importance" and "counterfactual reasoning" is not merely academic; it dictates the operational viability of AI in regulated industries.

3.3.1 Robustness and Stability Metrics

A critical yet often overlooked aspect of XAI is the robustness of the explanation itself. In our review of the literature, particularly the work surrounding LIME and SHAP [6, 12], a disturbing pattern emerges regarding the vulnerability of explanations to adversarial perturbations. It is possible to construct adversarial examples—inputs intentionally designed to cause the model to make a mistake—that also fool the explanation method.

For example, a "scaffolded" attack can allow a model to maintain its prediction accuracy while generating a completely innocent-looking explanation map. In a healthcare fraud detection scenario, an adversary could theoretically manipulate claims data to trigger a payment while the XAI module generates an explanation citing "standard compliance checks passed." This highlights a severe limitation of post-hoc methods: because they are approximations of the model and not the model itself, they can be decoupled from the true decision boundary.

Comparatively, contrastive methods that search for the "closest possible world" where the decision flips are inherently more robust because they directly interrogate the decision boundary rather than approximating it with a local linear model. By forcing the system to identify the Pertinent Negative, we are effectively stress-testing the model's decision logic. If a model claims a patient has pneumonia, and the contrastive explanation says, "It would be healthy if we removed this specific bone shadow," the radiologist immediately knows the model is hallucinating or focusing on artifacts. Thus, contrastive explanations serve a dual purpose: they explain the decision to the user, and they act as a debugging tool for the developer, revealing spurious correlations that feature-additive methods might obscure under a wash of "importance scores."

3.3.2 The Semantic Gap in Medical Imaging

When applying these concepts to medical imaging, the definition of "features" becomes problematic. Tabular data (age, blood pressure) has semantic meaning. Pixels do not. A SHAP heatmap overlay on an X-ray might highlight a region of the lung. To the algorithm, this is a cluster of pixel intensities. To the doctor, this must correspond to a biological tissue or pathology.

Current research indicates a "Semantic Gap" where XAI methods accurately point to where the model is looking, but fail to explain what it is seeing. A heatmap might light up a clavicle in a chest X-ray. Is the model looking at the bone density? The shape? Or is it detecting a bias artifact (like a hospital tag) often located near the clavicle?

Here, the "Glass Box" ideal faces its stiffest challenge. Purely pixel-based explanations (saliency maps) are often likened to a Rorschach test—users project their own knowledge onto the heatmap, leading to confirmation bias. If the AI highlights a vague area and the doctor suspects cancer, the doctor might interpret the highlight as confirmation of cancer, even if the AI was actually reacting to image noise.

To mitigate this, recent advancements propose "Concept Bottleneck Models" (CBMs). These models force the neural network to first predict high-level clinical concepts (e.g., "bone spurs," "lung opacity," "cardiomegaly") and then make a final diagnosis based only on those concepts. This creates a true Glass Box architecture. The explanation becomes: "Diagnosis is Pneumonia BECAUSE Lung Opacity is High AND Cardiomegaly is False." This is far superior to a SHAP heatmap because it operates on the level of medical ontology. However, CBMs require expensive, dense annotations at the concept level, which are rarely available in large quantities.

3.3.3 Business Implications: The Cost of Transparency

Moving to the business sector, specifically Fintech and Insurtech, the "Glass Box" requirement creates a tension between transparency and intellectual property (IP). As noted in the survey by

Shankheshwaria and Patel [11], businesses are hesitant to reveal the exact logic of their proprietary algorithms. A fully transparent Glass Box model essentially gives away the "secret sauce" of a credit risk model.

This creates a unique market for "Safe XAI"—explanations that satisfy regulatory requirements (like the Equal Credit Opportunity Act in the US or GDPR in the EU) without revealing the precise model architecture. Here, counterfactuals offer a strategic advantage. By telling a customer "You need to increase your savings by 10%," the bank provides a helpful, compliant explanation without revealing the exact weighting of "savings" in their internal scoring equation.

Furthermore, we must consider the liability introduced by XAI. If an AI system provides a specific explanation for a medical treatment recommendation, and that treatment fails, is the explanation itself a binding warranty? If a "Glass Box" model clearly shows it erred due to a specific missed feature, the path to malpractice litigation is straightforward. Conversely, if a "Black Box" model errs, the liability is murkier. This legal landscape suggests that while technical researchers push for maximum transparency, corporate legal teams may prefer a degree of opacity, or at least, carefully curated explanations that limit liability exposure.

3.3.4 Cognitive Load and the "Human-in-the-Loop"

Finally, we must address the human operator's capacity to process explanations. Jung et al. [3] discuss the "effectiveness" of explanations, but effectiveness is relative to the user's fatigue and expertise. In a high-volume environment—such as a radiologist reviewing hundreds of scans or a fraud analyst reviewing thousands of transactions—complex explanations are ignored.

A "Glass Box" that vomits distinct attribution scores for 50 different variables is effectively a Black Box to a tired human brain. This phenomenon, known as "Information/Choice Overload," suggests that XAI systems need a hierarchy of detail.

1. Level 1 (Triage): A simple traffic light (Red/Green) with a single-sentence rationale. (e.g., "High Risk due to cardiac history").

2. Level 2 (Review): A contrastive explanation for verification. (e.g., "Risk is high. If cardiac history were absent, risk would be low").

3. Level 3 (Audit): Full feature attribution (SHAP values) for deep investigation if the user disagrees with the initial assessment.

This tiered approach ensures that transparency mechanisms aid the workflow rather than obstructing it. It acknowledges that the goal of XAI is not just to be "correct," but to be "useful."

Discussion

The transition from "Black Box" to "Glass Box" systems is often framed as a technical hurdle—a need for better algorithms and more efficient computation. However, our comparative analysis suggests that the primary challenges are cognitive and systemic.

4.1 The Illusion of Transparency

There is a distinct danger in conflating "interpretability" with "truth." As methods like LIME and SHAP become ubiquitous, there is a risk that stakeholders will accept the explanation as a proxy for the mechanism. However, as we have discussed, post-hoc explanations are merely approximations. A LIME model might provide a perfect linear explanation for a decision that was actually made via a highly non-linear, chaotic process in a neural network. This creates an "Illusion of Transparency" where the user feels they understand the model, but their understanding is based on a simplified map, not the territory itself [18].

4.2 The Audience Problem

Gerlings et al. pose the critical question: "Explainable to whom?" [7]. The failure of many current XAI implementations stems from a one-size-fits-all approach. A developer debugging a model needs to see which neurons fired; they need feature importance. A patient needs reassurance; they need to know that their specific symptoms were considered. A regulator needs proof of non-discrimination; they need global statistical analysis.

We argue that the field must move away from generic "Explainability" metrics and toward role-specific metrics. A "Glass Box" for a doctor looks very different from a "Glass Box" for a data scientist. The former requires semantic alignment with medical training; the latter requires mathematical fidelity.

4.3 Trust vs. Reliance

Ultimately, the goal of XAI is to foster appropriate reliance. We do not want users to blindly trust AI (automation bias), nor do we want them to reject it out

of hand (algorithm aversion). Contrastive explanations appear to be the most effective tool for calibrating this trust. By showing the user the boundaries of the decision (e.g., "If the patient were 5 years younger, the risk would still be High"), the system allows the user to gauge the model's sensitivity. This builds a more robust form of trust—one based on an understanding of the model's limits, not just its successes.

4.Conclusion

In conclusion, while feature-additive methods like SHAP and LIME have laid the groundwork for interpretability, the future of high-stakes AI lies in contrastive and counterfactual reasoning. These methods align more closely with human cognitive processes, offering actionable insights and robust verification mechanisms. However, achieving the "Glass Box" ideal requires more than just new algorithms; it requires a fundamental shift in how we design AI interactions, prioritizing semantic meaning over mathematical abstraction and tailoring explanations to the specific needs of the human in the loop. Only by bridging this cognitive gap can we ensure that the powerful AI systems of tomorrow are not just accurate, but accountable.

5. References

1. Sheu, R.-K.; Pardeshi, M.S. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. Sensors 2022, 22, 8068.

2. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Trans. Neural Netw. Learn. Syst. 2020, 32, 4793–4813.

3. Jung, J.; Lee, H.; Jung, H.; Kim, H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon 2023, 9, e16110.

4. [Rai, A. Explainable AI: From black box to glass box. J. Acad. Mark. Sci. 2020, 48, 137–141.

5. Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access 2019, 7, 154096–154113.

6. Ribeiro, M.T.; Singh, S.; Guestrin, C. 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, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.

7. Gerlings, J.; Jensen, M.S.; Shollo, A. Explainable AI, but explainable to whom? An exploratory case study of xAI in healthcare. In Handbook of Artificial Intelligence in Healthcare: Practicalities and Prospects; Lim, C.-P., Chen, Y.-W., Vaidya, A., Mahorkar, C., Jain, L.C., Eds.; Springer International Publishing: Cham, Switzeralnd, 2022; Volume 2, pp. 169–198.

8. Dhurandhar, A., Chen, P.Y., Luss, R., Tu, C.C., Ting, P., Shanmugam, K., and Das, P. 2018. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In Neural Information Processing Systems.

9. Došilović, F. K., Brčić, M., and Hlupić, N. 2018. Explainable artificial intelligence: a survey. In 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, 0210–0215.

10. Ferreira, A., Madeira, S. C., Gromicho, M., Carvalho, M. d., Vinga, S., and Carvalho, A. M. 2021. Predictive medicine using interpretable recurrent neural networks. In International Conference on Pattern Recognition.

11. Yashika Vipulbhai Shankheshwaria, & Dip Bharatbhai Patel. (2025). Explainable AI in Machine Learning: Building Transparent Models for Business Applications. Frontiers in Emerging Artificial Intelligence and Machine Learning, 2(08), 08–15.

12. Gramegna, A. and Giudici, P. 2021. Shap and lime: an evaluation of discriminative power in credit risk. In Frontiers in Artificial Intelligence.