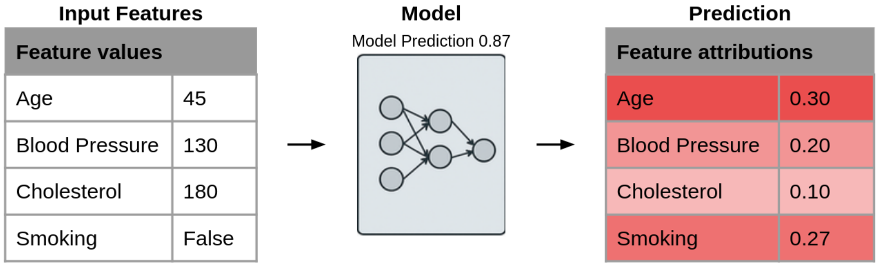
Local Causal Explainability of Machine Learning Models: Game-Theoretic Methods for Reliable Feature Attribution



Feature values	
Age	45
Blood Pressure	130
Cholesterol	180
Smoking	False

Feature attributions	
Age	0.30
Blood Pressure	0.20
Cholesterol	0.10
Smoking	0.27

Figure 1: Feature attribution pipeline

Introduction Artificial intelligence systems are increasingly making critical decisions that affect our daily lives—from medical diagnoses and loan approvals to criminal justice recommendations. While these AI systems can be highly accurate, they often operate as ”black boxes,” making it impossible for doctors, judges, or other professionals to understand why a particular decision was made. This lack of transparency becomes dangerous when AI systems make incorrect or biased decisions, potentially harming individuals and communities.

Current explanation methods, while popular, have fundamental flaws: they can provide misleading explanations that suggest false relationships between factors, or explanations that change unpredictably when the AI system learns from new data. For professionals who need to trust and verify AI decisions in high-stakes situations, these unreliable explanations can lead to poor decision-making or complete rejection of potentially beneficial AI tools.

Research Goals and Motivation This project aims to develop a new approach to AI explainability that provides causally consistent and locally reliable explanations. Instead of showing which factors were statistically associated with a decision, our method will take into the account the actual causal relationships that drive AI predictions—answering the crucial question of ”why” rather than just ”what.” Why this matters for human understanding: Humans naturally think in terms of cause and effect. When a doctor asks ”Why did the AI recommend this treatment?” they expect causal explanations like ”Because the patient’s symptoms indicate this condition” rather than statistical correlations. The conducted research recognizes that effective AI explanations must align with this fundamental aspect of human cognition. Our innovation combines two powerful mathematical frameworks:

- Game theory: To fairly distribute credit among different factors contributing to a decision.
- Causal discovery: To identify genuine cause-and-effect relationships rather than mere correlations.

The motivation is clear: as AI systems become more prevalent in healthcare, finance, and criminal justice, we need explanation methods that match how humans naturally reason about cause and effect, while remaining stable and trustworthy across different scenarios.

Expected Impact This research will produce theoretically grounded methods that provide mathematically guaranteed causal consistency, ensuring that AI explanations incorporate genuine cause-and-effect relationships rather than spurious correlations. These advances will translate into practical tools that work with real-world AI systems in high-dimensional settings, making the benefits accessible across diverse applications from medical diagnosis to financial decision-making.

Most importantly, this work will foster improved human-AI collaboration in critical domains where explanation quality directly impacts lives, allowing professionals to confidently leverage AI assistance while maintaining full accountability for their decisions.