



EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR RELIABLE DECISION SUPPORT

NAGARAJA K V

Associate Professor, Department of Computer Science Government First Grade Women's College, Tumkur
Junior College Field, B.H. Road, Tumkur, Karnataka 572102, India
Email: kvn.gsct@gmail.com

ABSTRACT

Machine learning predictions are increasingly accompanied by explanations intended to support human judgment. The usefulness of an explanation, however, depends on its faithfulness to the model, the assumptions used to generate it, and the task performed by its audience. This narrative review examines local surrogate explanations, additive feature attribution, counterfactual explanations, interpretable models, and selected methods for evaluating explanation reliability. Six foundational publications provide the basis for an analytical comparison. The paper develops a conceptual evaluation framework for educational decision support that separates predictive quality, explanation fidelity, stability, and human decision performance. It argues that an understandable explanation cannot establish the correctness, fairness, or causal validity of a prediction. A proposed pilot compares an interpretable baseline with a more complex model and evaluates whether explanations help users detect errors and respond appropriately to uncertainty. No new dataset, experiment, or measured improvement is reported. The principal contribution is a practical account of what should be tested before explanations are used to support consequential institutional decisions.

KEYWORDS: explainable artificial intelligence; interpretability; machine learning; feature attribution; counterfactual explanations; decision support.

1 INTRODUCTION

A predictive system can produce a score without providing an account that its user can examine. This creates a practical difficulty when a person must judge whether to act on that score, correct an input, or request further review. Explainable artificial intelligence addresses parts of this difficulty through descriptions of model behavior. Its purpose must be stated carefully because describing how a model responds to inputs is different from establishing why an outcome occurs in the world.

This paper asks how explanation methods can be evaluated for reliable decision support. It uses educational analytics as an illustrative setting in which staff might review a predicted need for student support. The setting is proposed for analysis and does not describe an implementation at the author's college. The central argument is that explanation quality requires several distinct tests; visual clarity or user confidence alone cannot demonstrate reliability.

2 REVIEW METHOD AND ANALYTICAL SCOPE

The paper uses a purposive narrative review of six primary research publications [1–6]. They were selected to represent widely discussed explanation approaches and direct critiques of explanation reliability. The selection includes local surrogate modeling, Shapley based attribution, counterfactual explanation, interpretable modeling, adversarial evaluation, and saliency sanity checks. The review is bounded and does not claim exhaustive coverage, systematic screening, or a quantitative synthesis of published results.

Each source is considered in relation to the question it answers, the information it requires, and the limits of its explanation. The resulting comparison informs a proposed evaluation protocol. Recommendations concerning educational deployment are the paper's conceptual synthesis rather than empirical findings from the cited studies. Published demonstrations are interpreted within their stated conditions and are not treated as universal performance guarantees.

3 WHAT AN EXPLANATION SHOULD EXPLAIN

A local explanation concerns a particular prediction, while a global account concerns broader model behavior. A feature attribution identifies how input variables contribute under a chosen explanation definition. A counterfactual describes an alternative input associated with a different model output. These accounts can all be informative, but they answer different questions. A user seeking an actionable change may receive little practical guidance from a ranking of variables that cannot be changed.

Interpretability can also be built into the predictive model. A small decision tree, a sparse linear model, or a constrained additive model can provide a structure that is directly inspectable, although its readability still depends on size, feature meaning, and the audience's knowledge. Rudin argues that interpretable models deserve priority in high stakes applications when they can provide adequate predictive performance [3]. This position motivates testing such models before assuming that complex predictors are necessary.

Consider an illustrative model that associates low attendance with a higher probability of noncompletion. An attribution to attendance describes the model's response; it does not establish that changing attendance alone would change the real outcome. Unobserved circumstances may influence both variables. The explanation interface should therefore avoid causal wording unless the study design supports a causal claim.

Explanations should also identify their scale. A positive contribution might increase a log odds score rather than a probability. A reference value might represent a selected background population rather than a neutral individual. Without this information, a technically correct attribution can be misunderstood. Defining the user's question and the output scale is consequently part of the explanation specification.

4 LOCAL SURROGATES AND FEATURE ATTRIBUTION

Ribeiro et al. introduce LIME, which fits an interpretable approximation around a prediction of interest [1]. The method samples perturbed inputs, obtains the original model's outputs, and fits a simpler local surrogate. Its appeal is that the predictive model can remain unchanged while the explanation uses a representation that a person can inspect. The account is local and should not be interpreted as a complete description of the predictor.

The surrogate depends on its neighborhood, proximity weighting, sampling procedure, and complexity constraint. Perturbations that combine incompatible input values can create an artificial neighborhood. Repeating an explanation with different random seeds or plausible neighborhood definitions can reveal sensitivity. Local agreement should be measured on a separate neighborhood sample; a compact list of influential variables is not itself evidence that the surrogate closely tracks the model.

Lundberg and Lee introduce SHAP as a framework for additive feature attribution [2]. Its explanations distribute a prediction relative to a reference value across input features under a specified attribution game. The formal properties apply within the framework's assumptions. They do not automatically establish that the explanation identifies a causal mechanism or that every approximation has identical numerical accuracy.

The reference population and treatment of feature dependence require explicit choices. With correlated variables, different definitions of a missing feature can distribute credit differently. A researcher should record the background data selection, the explainer implementation, and the output being explained. Averaging absolute attribution values across records can summarize model dependence, but that summary loses the direction and case specific context of individual contributions.

5 COUNTERFACTUAL EXPLANATIONS AND FEASIBILITY

Wachter et al. examine counterfactual explanations that describe changes associated with a desired prediction without requiring disclosure of the entire model [4]. Such an explanation can be expressed as a search for an alternative input x' that changes the predicted class while remaining close to the observed input x . The distance function determines what counts as a small change and therefore has substantive consequences.

For a proposed educational application, a counterfactual should distinguish fixed attributes, variables a student can influence, and variables that require institutional action. A numerically nearby record may be impossible because it violates time order or relationships between fields. Counterfactual validity at a model threshold consequently does not establish feasibility, affordability, or a beneficial real world intervention. These properties require additional evidence and explicit constraints.

6 RELIABILITY RISKS IN EXPLANATION SYSTEMS

An explanation can remain readable even when it provides a weak account of the predictor. Slack et al. demonstrate attacks against LIME and SHAP in which a predictor can behave differently on ordinary inputs and on perturbations associated with explanation procedures [5]. This shows a vulnerability under the studied attack conditions. It motivates evaluating the explanation pipeline against deliberate manipulation rather than assuming that model agnostic access guarantees transparency.

Adebayo et al. propose sanity checks for saliency methods that examine sensitivity to changes in learned

parameters and training labels [6]. The central lesson for evaluation is that an attractive visualization may not demonstrate dependence on what the model learned. Although saliency maps concern a different representation from tabular attribution, the general requirement to test the source of an explanation remains relevant. Each explanation family still needs tests appropriate to its own construction.

Stability is another concern. When inputs change slightly without a meaningful change in the prediction, a radically different explanation can confuse users. However, instability must be interpreted in context: a model near a genuine decision boundary may have changing local behavior. Evaluation should distinguish arbitrary variation caused by the explanation procedure from variation that reflects the predictor itself.

Explanations may also increase inappropriate confidence. A user could accept a wrong recommendation because its supporting chart appears persuasive. A reliable interface should help people identify both useful and unreliable predictions. User satisfaction and perceived transparency therefore require separate measurement from error detection, decision accuracy, and willingness to seek review. These are proposed evaluation priorities, not claims of observed effects in the present paper.

7 PROPOSED FRAMEWORK FOR EDUCATIONAL DECISION SUPPORT

The proposed framework begins with an operational question: what decision will a staff member make after seeing the prediction? A narrowly defined example is whether to offer a voluntary support discussion. The pilot would exclude automated sanctions or denial of opportunities. A documented intended use helps determine which input variables are relevant and which explanatory statements would be misleading.

The development team would build an interpretable baseline and a more complex candidate using the same eligible inputs and outcome definition. It would then compare their predictive performance before adding explanation interfaces. A complex model would require a demonstrated benefit that is meaningful for the intended task. Explanation complexity, runtime, and maintenance would be counted among its costs.

7 PROPOSED FRAMEWORK CONTINUED

The data preparation stage would record variable definitions, collection times, missing value rules, and the relationship between repeated records. Variables measured after the predicted outcome would be excluded. Student identifiers would serve partitioning and linkage purposes where authorized, rather than becoming predictive inputs. Sensitive attributes, if used for approved evaluation, would require restricted handling and careful reporting. Removing an attribute from training does not demonstrate that correlated proxies are absent.

The interface would display the model version, input date, predicted outcome definition, and any uncertainty information supported by the evaluation. A local explanation would state its reference population and limitations in ordinary language. Staff should be able to identify an incorrect input, record disagreement, and request further assessment. Explanation outputs should be stored only to the extent needed for review, with access and retention matched to their potentially sensitive content.

8 PROPOSED EVALUATION PROTOCOL

The first evaluation layer concerns prediction. Data partitions should reflect the intended use, with time based testing when the objective concerns later cohorts. All records for a student should remain within a single partition

where leakage would otherwise occur. Preprocessing and hyperparameter choices would be fitted without access to the final test set. Metrics should include discrimination, calibration, and threshold dependent errors relevant to the proposed support workflow.

The second layer concerns explanation fidelity. For local surrogates, researchers would measure agreement with the predictor on held out, plausible neighborhood samples. For additive attributions, they would document the reference data and check reconstruction of the explained output within numerical tolerance where the method promises that property. Exact reconstruction is an internal consistency check; it does not establish that the reference or feature dependence assumptions are appropriate.

The third layer evaluates stability and counterfactual feasibility. Repeated runs, alternative background samples, and small plausible input changes would test attribution sensitivity. Counterfactuals would be checked for target attainment, immutable feature constraints, valid feature combinations, and persistence under modest model changes. Researchers would report failure rates as well as successful examples. No counterfactual would be described as guaranteed real world recourse solely because it changes a model prediction.

The final layer evaluates human use. A preregistered study could compare prediction only, interpretable model output, and prediction with a post hoc explanation. Participants would review cases with known test outcomes, including intentionally selected model errors. Appropriate reliance, error detection, review requests, and task completion time would be measured alongside comprehension. The protocol is proposed; no participants have been recruited and no outcomes are reported here.

9 DISCUSSION AND RESEARCH IMPLICATIONS

The framework treats reliability as a property of the full decision support process. A predictor can be accurate while its explanation is unstable, and an explanation can faithfully describe a predictor that is poorly calibrated. Separating these properties prevents a strong result on one measure from masking a failure on another. A deployment report should therefore present predictive performance, explanation tests, and user study findings as distinct bodies of evidence.

The practical value of an explanation also depends on available actions. If staff cannot correct inputs, seek more information, or offer an appropriate response, an attribution chart may add little operational value. For an educational pilot, the explanation should be assessed against a defined support procedure. The research question then becomes whether the information improves a relevant judgment, rather than whether users simply prefer a more elaborate interface.

Fairness assessment remains necessary even when a model is understandable. An interpretable rule can encode a harmful relationship, while a post hoc explanation can reveal associations without establishing whether the decision process is acceptable. A proposed pilot should examine error patterns across relevant groups when authorized data and adequate sample sizes exist. Aggregate summaries should avoid exposing individuals or presenting unstable small group estimates as precise findings.

A deployment decision should use predefined acceptance criteria. These might require acceptable calibration, no substantial deterioration for an evaluated subgroup, adequate explanation fidelity under plausible perturbations, and improved detection of incorrect recommendations. Numerical thresholds must reflect the application and be fixed before final testing. This paper proposes the structure of such criteria without inventing values or claiming that any existing system has satisfied them.

10 LIMITATIONS AND CONCLUSION

The analysis is limited to six foundational sources and does not cover the full development of explainability research. It includes no benchmark dataset, software comparison, institutional field study, or causal analysis. The educational setting and evaluation framework are conceptual. Their suitability depends on the prediction task, the resources available, the quality of local records, and the people who would use the system.

Explainable artificial intelligence can support examination of model behavior when its purpose and assumptions are made explicit. Reliable use requires more than an understandable presentation. Researchers should compare interpretable baselines, test the fidelity and stability of explanations, constrain counterfactuals to plausible alternatives, and measure whether people recognize errors. The next step is a small reproducible pilot that evaluates these requirements together. Until that evidence exists, explanations should be treated as information for critical review rather than proof of correctness or justified trust.

REFERENCES

- [1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,
- [2] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. Author manuscript: <https://arxiv.org/abs/1705.07874>
- [3] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence.
- [4] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. arXiv preprint arXiv:1711.00399. <https://arxiv.org/abs/1711.00399>
- [5] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.
- [6] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31. Author manuscript: <https://arxiv.org/abs/1810.03292>