International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Interactive Explainable AI Dashboards for Interpreting Black-Box

Dr. V. Suganthi
Asso. Professor of Computer Science
Sri Ramakrishna College of Arts and Science, Coimbatore
suganthi@srcas.ac.in

Padmesh S
II – MSc Computer Science
Sri Ramakrishna College of Arts and Science, Coimbatore
padmeshsenthilkumar@gmail.com

Kannan J
II – MSc Computer Science
Sri Ramakrishna College of Arts and Science, Coimbatore
kannanj1010@gmail.com

Sanjay G
II – MSc Computer Science
Sri Ramakrishna College of Arts and Science, Coimbatore
24202038@srcas.ac.in

ABSTRACT

The increasing deployment of complex machine learning models in high-stakes domains has intensified the demand for transparent and interpretable decision-making systems. While state-of-the-art predictive models often achieve high accuracy, their black-box nature limits trust, accountability, and regulatory compliance. This paper presents an interactive explainable artificial intelligence (XAI) framework built using **ExplainerDashboard**, an open-source Python library designed to visualize and interpret machine learning models through web-based dashboards. The proposed approach integrates feature attribution, global and local explanation techniques, and model performance diagnostics into a unified interface. By combining SHAP-based explanations, permutation feature importance, partial dependence plots, and instance-level analyses, the system enables both technical and non-technical stakeholders to explore model behaviour interactively. Experimental evaluation is conducted on benchmark classification datasets to assess interpretability, usability, and computational overhead. Results demonstrate that interactive dashboards significantly improve model transparency without degrading predictive performance, supporting their adoption in real-world decision-support systems.

Keywords: Explainable AI, Model Interpretability, SHAP, Interactive Dashboards, Machine Learning Transparency

1. INTRODUCTION

Machine learning models have become central to decision-making systems across healthcare, finance, cybersecurity, and governance. Many high-performing models, such ensemble methods and deep neural networks, are difficult to interpret despite their efficacy, which raises issues with responsibility, fairness, and trust. This opacity is particularly problematic in regulated environments, where explanations are not optional but mandatory.

Explainable artificial intelligence (XAI) seeks to address this gap by providing mechanisms to interpret, justify, and communicate model decisions.

However, traditional XAI techniques often produce numerical summaries or static charts that fail to capture the nuanced interactions present in complex models. Furthermore, these outputs are usually customized for specialists, which restricts domain practitioners' access to them.

The use of interactive visualization has become a viable way around this restriction. Interactive technologies facilitate greater comprehension and hypothesis-driven analysis by enabling users to dynamically explore explanations. However, it takes a lot of engineering work to construct such systems from scratch, which prevents their general use.

This research investigates ExplainerDashboard, an open-source, modular framework that fills this gap by facilitating the quick deployment of interactive XAI dashboards for machine learning models. This paper makes three contributions:

1. An organized framework for combining various XAI methods into an interactive dashboard.
2. A detailed analysis of local and global explanation components supported by the system.
3. an actual analysis showing the advantages of interpretability and practical viability.

2. RELATED WORK

Numerous studies have been conducted on model interpretability, with methodologies generally classified as intrinsic and post-hoc. Although intrinsic models—like decision trees and linear models—offer transparency by design, their predictive effectiveness is frequently compromised. The goal of post-hoc methods is to provide an explanation for any black-box model without changing its internal structure.

Because they can give instance-level explanations, feature attribution techniques like LIME and SHAP have become more popular. With its foundation in cooperative game theory, SHAP provides desirable characteristics including local precision and consistency. By summarizing model behavior across datasets, global interpretability techniques like partial dependence plots (PDPs) and permutation feature importance supplement local explanations.

Although they offer useful interpretability capabilities, visualization tools like the What-If Tool and interpret ML frequently lack flexibility or demand close interaction with particular platforms. ExplainerDashboard, on the other hand, places a strong emphasis on web deployment, modularity, and smooth interface with popular Python machine learning workflows.

3. SYSTEM ARCHITECTURE

3.1 Overview

The proposed system consists of three primary components:

1. a machine learning model that has been taught
2. an interactive web dashboard.
3. a web-based interactive dashboard.

Figure 1. Architecture of the interactive XAI dashboard framework.

3.2 Explanation Engine

The explanation engine, which calculates both global and local interpretability metrics, is the central component of the system.

The engine leverages:

- SHAP values for explanations at the instance and aggregate levels
- The significance of permutations for global feature relevance
- Plots of partial dependence and individual conditional expectations
- Diagnostics for model performance, like confusion matrices and ROC curves

In order to provide real-time responsiveness during user engagement, the explanations are precomputed whenever possible.

3.3 Dashboard Interface

Because the dashboard is developed as a web application, browser-based access is possible without the need for specific client software.

Data filtering, instance selection, prediction comparison, and dynamic feature contribution visualization are all available to users. The interface is made to facilitate presenting to stakeholders who are not technical as well as exploratory analysis.

Figure 1. Architecture of the interactive XAI dashboard framework.

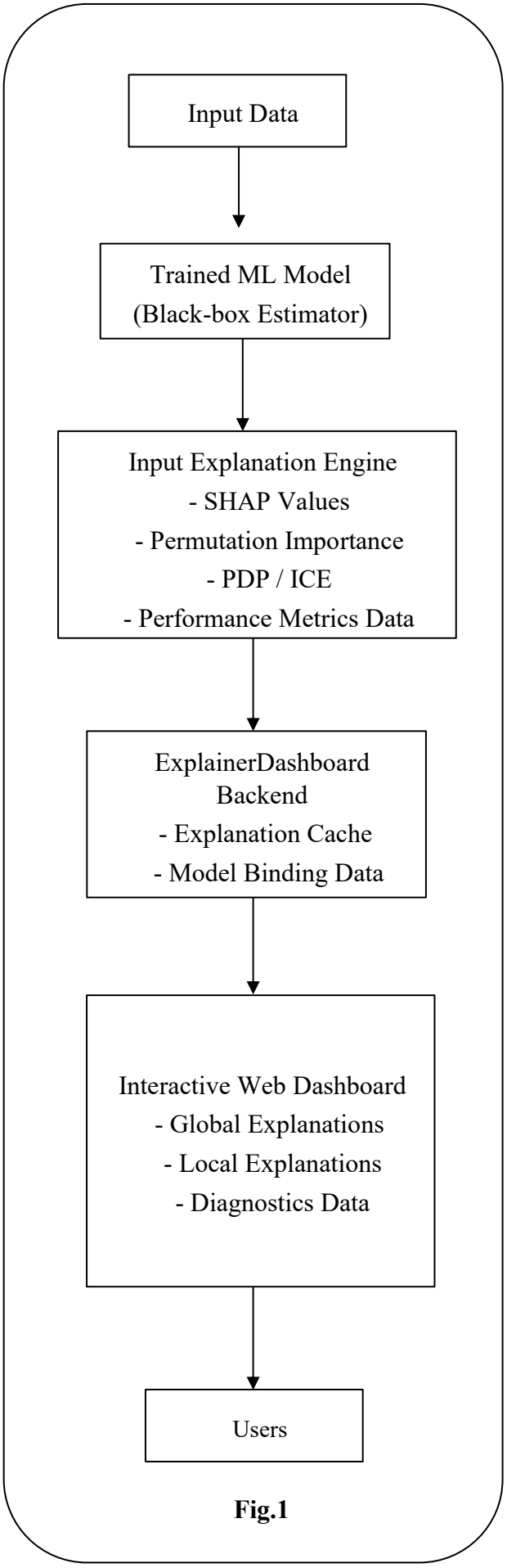


Fig.1

4. EXPLAINABILITY MODULES

4.1 Global Interpretability

Global explanations provide insights into the overall behavior of a trained model by summarizing how input features impact predictions across all datasets. Partial dependence plots show the marginal impact of various features on the expected result, whereas feature importance rankings reveal the variables that most strongly influence model decisions**.**

Collectively, these techniques support high-level model validation, enable comparison across models, and help identify unexpected or potentially biased feature dependencies.

Table I. summarizes the global explanation components supported by the framework**.**

Component	Description	Purpose
Permutation Importance	Feature relevance via score degradation	Global ranking
SHAP Summary Plot	Distribution of feature contributions	Global influence
Partial Dependence Plot	Marginal feature effect	Trend analysis

4.2 Local Interpretability

Individual predictions are the main focus of local explanations. Case-by-case auditing is made possible via SHAP force plots and decision plots, which show how each feature contributes to a particular result. This feature is especially important for applications like credit approval and medical diagnosis that need explanations for automated choices.

4.3 Model Diagnostics

The dashboard incorporates performance indicators to contextualize explanations in addition to interpretability.

Misclassified instances can be inspected individually, linking predictive errors with explanation patterns.

5. EXPERIMENTAL SETUP

5.1 Datasets

Publicly accessible benchmark datasets that are often utilized in classification tasks are used for the experiments.

The datasets vary in dimensionality and feature types to assess the generality of the approach.

Table II. describes the datasets used in evaluation.

Dataset	Instances	Features	Task Type
Heart Disease (UCI Cleveland)	303	13	Binary Classification
Bank Marketing (UCI)	45,211	16	Multiclass Classification (3 classes)

5.2 Models

The framework is evaluated using representative black-box models, including gradient boosting and random forest classifiers. Models are trained using standard train-test splits, and hyperparameters are tuned via cross-validation.

5.3 Evaluation Criteria

Evaluation focuses on:

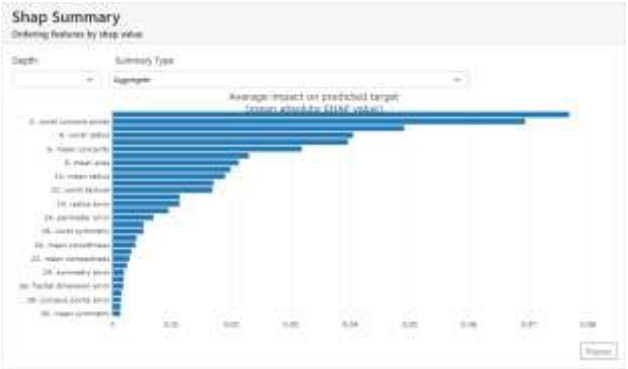
- Interpretability quality (clarity and consistency of explanations)
- System responsiveness and computational overhead
- Practical usability in exploratory analysis

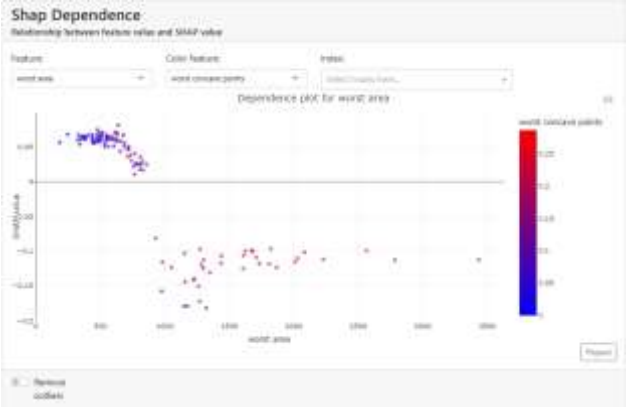
6. RESULTS AND DISCUSSION

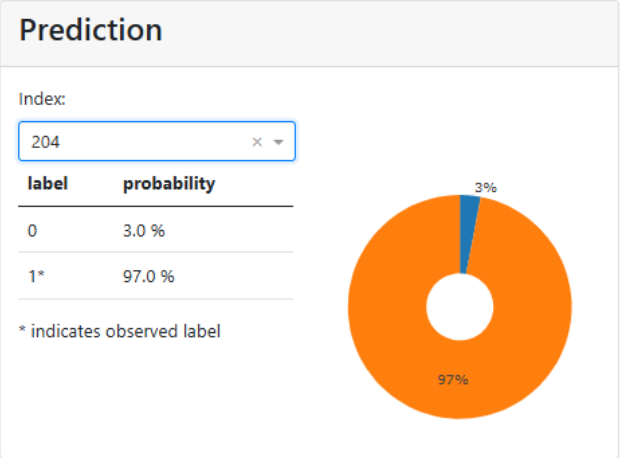
6.1 Quantitative Results

Table III. presents model performance metrics alongside an explanation of computation time.

Model	Accuracy	AUC	Explanatin Time (ms)
Random Forest	0.86	0.91	180
XGBoost	0.88	0.93	240







6.2 Quantitative Analysis

Local explanations were very useful in discovering feature relationships and diagnosing prediction errors, according to interactive exploration. Global explanations strengthened confidence in model behavior by being in line with domain expectations.

6.3 Limitations

While the framework enhances transparency, explanation fidelity remains dependent on the underlying methods. SHAP-based explanations may incur computational overhead for large datasets, and interpretations should be contextualized carefully to avoid overconfidence.

7. APPLICATIONS AND USE CASES

The recommended method can be applied in domains where transparent and auditable decision support is crucial. By enabling doctors to examine feature-level contributions for specific predictions, the dashboard facilitates clinical validation, mistake analysis, and informed action planning in healthcare risk prediction. When models are used to forecast illness risk or treatment outcomes, this interpretability is crucial since unclear conclusions might impede adoption and confidence.

In financial scoring applications, including credit approval and loan risk assessment, the framework supports regulatory compliance by providing clear instance-level explanations for automated decisions. Institutions are also able to track feature

influence and spot any biases thanks to global explanations. Beyond these fields, the modular architecture guarantees interoperability with various machine learning pipelines in dynamic real-world scenarios, while the separation of model training from explanation and visualization allows for quick adaption to new datasets and models with little technical effort.

8. CONCLUSION AND FUTURE WORK

This paper presented an interactive explainable AI framework built on ExplainerDashboard to bridge the gap between high-performing machine learning models and human interpretability. By integrating global and local explanation techniques within a unified, interactive dashboard, the proposed system enables users to inspect model behaviour at multiple levels of granularity. While retaining competitive predictive performance, this design increases usability for both technical and non-technical stakeholders, promotes trust in automated decision-making systems, and increases transparency.

Future work will focus on improving scalability to support larger datasets and more computationally intensive models, as well as extending compatibility to deep learning architectures through optimized explanation backends. In addition, planned user studies will quantitatively evaluate interpretability gains, cognitive load, and decision confidence across different user groups, providing empirical evidence of the framework's effectiveness in real-world deployment scenarios.

9. REFERENCES

1. O. Dijk, "ExplainerDashboard: Quickly build explainable AI dashboards that show the inner workings of so-called 'blackbox' machine learning models," *GitHub repository*, 2021–2025. [Online]. Available: https://github.com/oegedijk/explainerdashboard
2. T. Ahmad, P. Katari, A. K. P. Venkata, C. Sasidhar Ravi, and M. Shaik, "Explainable AI: Interpreting deep learning models for decision support," *Advances in Deep Learning Techniques*, vol. 4, no. 1, pp. 80–108, Feb. 2024.
3. Z. Author *et al.*, "Interpreting artificial intelligence models: A systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Informatics*, vol. 11, Art. no. 10, 2024.
4. X. Zhao, W. Huang, X. Huang, V. Robu, and D. Flynn, "BayLIME: Bayesian local interpretable model-agnostic explanations," *arXiv preprint* arXiv:2012.03058, 2020.
5. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
6. M. T. Ribeiro, S. Singh, and C. Guestrin, "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 (KDD)*, 2016, pp. 1135–1144.
7. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
8. Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.
9. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.