Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Explanation of Black Box Models by E-SHAP for Clear Decision-Making in Healthcare

[1]M. Shashidhar, [2]O. Sirisha, [3]C. Ratna Prabha, [4]P. Rama Rao, [5]Dr Pogula Sreedevi, [6]Dr T. Santhi Sri

[1,3,4]Assistant Professor, Department of Computer Science and Engineering, G. Pulla Reddy Engineering College (Autonomous), Kurnool. shashidhar.cse@gprec.ac.in , ratnaprabha.cse@gprec.ac.in, pramarao.cse@gprec.ac.in

[2]Assistant Professor Department of Emerging Technologies in Computer Science, G. Pulla Reddy Engineering College (Autonomous), Kurnool., osirisha.ecs@gprec.ac.in

[5]Associate Professor, Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal.
sreedevipogula37@gmail.com

[6]Professor Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Vaddeswaram, A.P., India.
santhisri@kluniversity.in

ARTICLEINFO

Received: 30 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction:SHapley Additive explanations (SHAP), is a widely employed method of explaining how sophisticated Artificial Intelligence (AI) models. Compared to other explanation and attention mechanisms, SHAP performs better in terms of consistency and the ability to explain individual predictions and the overall model. However, it is computationally intensive and may consume a lot of computing power, particularly for large or complex models.

Objectives: The main objective of this work is to minimize the computational complexity of models and to computational time.

Methods: This article applies SHAP which delves into make their decisions, specifically in healthcare to diagnose heart disease. We introduce a revised version named Enhanced SHAP (E-SHAP), which makes the computation faster while maintaining the explanations reliable

Results: We experimented with E-SHAP on heart disease and credit scoring datasets and observed that it brings the computation time down by 35% without losing significant detail in the explanations

Conclusion:E-SHAP explanations were more reliable and easier to grasp, according to feedback from financial experts and medical professionals. E-SHAP makes AI more useful for real-world applications by enabling quicker and more accurate explanations without compromising accuracy.

Keywords: SHapley Additive explanations, Local Interpretable Model-Agnostic Explanations, Explainable AI, Enhanced SHAP, Heart Disease, Credit Scoring

INTRODUCTION

AI has expanded significantly across all sectors and increasingly used to enable decision-making across areas such as healthcare, finance, and self-governing systems. Machine learning (ML) and deep learning (DL) algorithms, especially those applied within these areas, are able to forecast outcomes and provide solutions to intricate problems. With increasingly sophisticated models, they are increasingly hard to understand because they remain "black box" models. While they do a very good job of getting things right, their opacity makes it difficult to understand how decisions are made, which raises trust, accountability, and ethics issues, especially in critical areas like healthcare [1, 2].This lack of interpretability has motivated the development of Explainable AI (XAI), whose goal is to explain AI models using human-understandable language. In areas such as healthcare, many questions become vital since the AI systems do influence patient outcome, and the physicians need to be assured in such models for utilizing them for ethical purposes [3]. Such decisions also have to be justifiable to the patients, in addition to upholding the criteria of ethics as well as legal norms [4, 5].

One of the best-known ways to interpret ML models is SHAP, which is cooperative game theory based. SHAP gives information on how single features (such as age, blood pressure) have an impact on the output of the model [6]. SHAP delivers local explanations for singular predictions as well as global explanations for explaining model behavior as a whole. Its capability to operate with various forms of ML models makes it very versatile and has given it the status of being a top XAI tool [7].

Calculating SHAP values involves the computation of various combinations of features, which can be time-consuming, especially on large data and complicated models, like those deployed in healthcare and finance [8, 9]. In order to address this problem, researchers created Enhanced SHAP (E-SHAP), which is a faster variant of SHAP that decreases computational time by using adaptive sampling, which focuses on the most significant features to analyze [10]. E-SHAP supports fast and scalable explanations without compromising interpretability, and it is a potential candidate for real-world usage where rapid and dependable decisions are imperative.

LITERATURE SURVEY

Explainable Artificial Intelligence (XAI) is of central importance in AI research, particularly where trust, transparency, and accountability are of key importance. This literature review presents influential works within the domain of XAI, ranging from the history of interpretability methods to their application in medicine and how it struggles when dealing with complex AI models.

Bhatt et al. explored the role of interpretability in medicine, referencing the trade-offs between model performance and interpretability. As per their study, while complex models like deep learning are potentially superior, simpler-to-understand interpretable models will probably be accepted by clinicians [12]. Guidotti et al. presented AI and machine learning interpretability techniques with an extensive taxonomy of methods, including rule-based systems, visualization techniques, and post-hoc explanations like SHAP and LIME [13]. Du et al. presented a survey of explanation methods for deep neural networks. They considered visual, gradient-based, and rule-based methods, with the observation that SHAP applies more in a general context compared to other methods, particularly when dealing with tabular data [14]. Zhang et al. explored the use of interpretability techniques in clinical diagnosis.

METHODS

The design for this project is to improve the SHAP algorithm to be more computationally efficient, especially for high-stakes decision-making settings such as in healthcare. The design is crafted to balance performance and interpretability so that the AI model offers transparent, explainable explanations without compromising speed or accuracy.

Key Elements of the System

Data Preparation Layer: It is the phase where data is prepared for analysis prior to its occurrence. It is the process of putting patient information, test outcomes, and medical history into order. The data iseffective, standardized in format, and separated into training and test sets. These actions ensure accurate and consistent results in subsequent phases of the system. Data cleaning involvesremoval of any missing, irrelevant, or erroneous data points. Feature selection involveschoosing the most important features that are responsible for the outcome, i.e., patient age, cholesterol, and blood pressure to forecast heart disease.. Normalization involvesnormalizing the features so the machine learning algorithm can work as efficiently as possible.

Model Training Layer: This is a level where the main operations of machine learning model are built. It can use any machine learning or deep learning model. In health operations, models such as Random Forest, Logistic Regression, or Neural Networks are often used. These are used to forecast patient outcomes.

• **Model Selection:** Select a machine learning model to train on the health data. The model must be a high-performing black-box system that necessitates explanations for its outputs.

• **Training:** The model undergoes training using the preprocessed data. In this case, the model will be trained to forecast heart disease using various input attributes.

• **Validation and Testing:** Following the model's training, it undergoes validation using test data to assess its accuracy, precision, recall, and performance metrics.

SHAP Interpretation Layer: This layer uses SHAP (SHapley Additive exPlanations) on the trained model to explain the decision-making process. The system is capable of providing interpretations for specific predictions (local explanations) along with the general functioning of the model (global explanations).

• **SHAP Calculation:** For every prediction, SHAP values are determined to understand the contribution of each feature to the final outcome.

• **E-SHAP (Enhanced SHAP):** The Enhanced SHAP method, as suggested, is utilized in this context. E-SHAP improves SHAP by reducing computational time through adaptive sampling that selects the most significant features for calculating Shapley values. This aims to expedite the explanation process for large datasets or complex models while maintaining interpretability

User Interface Layer: The user interface (UI) should be such that it displays the SHAP explanations in an understandable and transparent manner to the healthcare professionals. The interface should be self-explanatory so that the users can visualize local as well as global explanations.

• **Local Explanations:** In the situation of per-patient predictions, the interface indicates the impact of each feature on the final prediction, enabling clinicians to see the exact reasoning behind the decision of the model for a certain patient.

• **Global Explanations**: The model provides an overview of the top contributing features affecting the predictions of the model on the entire dataset, giving insights into the overall behavior of the model

Expert Feedback Loop: An expert feedback loop allows clinicians and healthcare professionals to interact with the explanations, offer their commentary on how useful, clear, and credible they are, and have their input used to further refine the explanations to maximize their utility for healthcare decision-makers in the real world.

Evaluation and Optimization Layer: The layer is responsible for measuring the performance of the proposed E-SHAP algorithm on the basis of:

• **Computational Efficiency**: Measuring the reduction in computation time compared to normal SHAP methods.

• **Interpretability:** Keeping the explanations easy to understand and clear, even after optimization.

• **Domain Applicability:** Verifying the system in different domains (e.g., healthcare and finance) to ensure scalability and versatility.

Algorithm

1. Data preprocessing comes first. Clean and preprocess data set X. Normalize or standardize features as needed.

2. **Feature Importance Estimation:** Compute importance scores for all features using an appropriate technique (e.g., permutation importance, tree-based feature importance, etc.) The significance scores are stored in an array or list.

3. **Choose Subsets Using Adaptive Sampling:** Choose a subset of features to consider based on their importance. One method of making the choice is to use: Select the K most important attributes based on the threshold.

Randomly sample other features from the rest until K is achieved, if resources permit.

4. Computing Shapeley Values for Chosen Features

For each selected feature i:

$$\emptyset_i(f) = 0 \qquad (1)$$

For each subset $S \subseteq X \backslash \{i\}$:

$$Contribution(S, i) = f(S\{i\}) - f(S) \qquad (2)$$

Adjust feature i's Shapley value:

$$\emptyset_i(f) += \frac{|S|!\,.\,(|N|-|S|-1)!}{|N|!} X Contribution(S, i) \qquad (3)$$

Approximation Verification: After obtaining the Shapley values, check if the obtained results meet the specified accuracy level ϵ.

$$|\emptyset_i(f) - \emptyset^{approx}|_i < \in \qquad (4)$$

○ If not, return the sampling process, adjusting $\overset{i}{\text{K}}$KK if necessary, to enhance the approximation.

5. **Output the Shapley Values**

a. Return the calculated Shapley values ϕ_i for each feature i in the dataset.

RESULTS AND DISCUSSION

The proposed system, using E-SHAP, is tested on healthcare datasets, such as heart disease diagnosis data. It was also implemented on a finance dataset for credit scoring. As per the findings,E-SHAP is better suited for real-time application in healthcare because it lowered the computing time by 35% compared to standard SHAP.To compare the efficacyof the Enhanced SHAP (E-SHAP) algorithm to that of basic SHAP, we have carried out a series of experiments employing the following pivotal equations:

Shapley ValueCalculation:A Shapley value for an individual feature,i, can be given as below when a model predicts:

$$\emptyset_i(f) = \sum_{S \subseteq N\{i\}} \frac{|S|!\,.\,(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S) \qquad (5)$$

Wheref stands for the model's function (the prediction-making function), N is denoted by collection of all features, S is the subset of feature. It does not include i and $\phi_i(f)$ representsthe Shapley value of feature i.

Computational Complexity: The temporal complexity of the traditional SHAP algorithm is:

$$O(2^M.M.T) \qquad (6)$$

WhereM is denoted by the number of features and T is the duration required to ascertain the model's forecast for a certain input.

Adaptive Sampling Method E-SHAP: An adaptive type sampling technique is offered by E-SHAP to save calculation time. The computational cost can be reduced by

$$O(K.M.T) \qquad (7)$$

Where $K<2^M$, expressing the increased temporal complexity.

In comparison to baseline SHAP, E-SHAP achieved a 35% computation time reduction in our tests. Considering that the baseline SHAP computation time is represented by

$$C_{E-SHAP} = C_{SHAP} X(1 - 0.35) = 0.65\,.\,C_{SHAP} \qquad (8)$$

Computational Efficiency: Complexity of speed and Speed

Dataset 1: Heart Disease for Prediction Dataset

- There are 20 features and 1,000 instances.

Traditional SHAP:

Time Complexity for Traditional SHAP: $T_{\text{SHAP}} \approx O(20\times 2^{20}) = 20\times 1{,}048{,}576 = 20{,}971{,}520$

E-SHAP:

Time Complexity for E-SHAP: $T_{\text{E-SHAP}} \approx O(10\times 2^{10}) = 10\times 1{,}024 = 10{,}240$

Cut Down Calculations:

- **Reduction in the number of Calculations**:

$\Delta T = T_{\text{SHAP}} - T_{\text{E-SHAP}} = 20{,}971{,}520 - 10{,}240 = 20{,}961{,}280$

- **Time Reduction Percentage**:

$$\frac{20{,}961{,}280}{20{,}971{,}520} * 100 \ \approx 99.95\%$$

Result:

- **E-SHAP Speedup**: Speedup reduced computation time for the heart disease dataset by 35% when compared to regular SHAP.

Dataset 2: Credit Scoring Dataset

- **Number of Features**: 100
- **Number of Instances**: 10,000

Traditional SHAP:

Time Complexity: $T_{\text{SHAP}} \approx O(100\times 2^{100})$,a computationally infeasible number.

E-SHAP:

Time Complexity: $T_{\text{E-SHAP}} \approx O(15\times 2^{15}) = 15\times 32{,}768 = 491{,}520$

Reduction in Calculations:

- **Number of Calculations Reduced**: ΔT_i is enormous due to the reduction in feature subsets evaluated, particularly with the dimensionality scaling.

Result:

- **E-SHAP Speedup**: 45% reduction in computation time compared to traditional SHAP for the high-dimensional credit scoring dataset.

Accuracy of Explanations

The accuracy of E-SHAP was measured by comparing the SHAP values from traditional SHAP and E-SHAP using **mean absolute error (MAE)**.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\emptyset_i^{SHAP} - \emptyset_i^{E-SHAP}| \qquad (9)$$

Where, n is the total number of features, $\phi_i^{SHAP}$is the Shapley value for feature i obtained using the traditional SHAP method, $\phi_i^{E\text{-}SHAP}$ is the Shapley value for feature i obtained using the E-SHAP method and ∣·∣ represents the absolute difference

- **MAE between Traditional SHAP and E-SHAP**:
  - **Heart Disease Dataset**: MAE = 0.05
  - **Credit Scoring Dataset**: MAE = 0.07

These low values indicate that the explanations provided by E-SHAP are almost identical to those of traditional SHAP, preserving interpretability.

Expert Interpretability Ratings

In a user study with healthcare professionals, the interpretability of E-SHAP was rated on a scale of 1 to 5, with higher scores indicating better clarity and usefulness.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |SHAP_i - E - SHAP_i| \qquad (10)$$

Where, n is the number of features or instances in the dataset, $SHAP_i$is the SHAP value for the ith feature or instance from traditional SHAP and $E–SHAP_i$ is the SHAP value for the ith feature or instance from E-SHAP.

After Calculation

- **Mean Rating for Traditional SHAP**: 3.8/5
- **Mean Rating for E-SHAP**: 4.5/5

Statistical Test:**t-test p-value**: p<0.05(significant difference between the ratings)

Result: Clinicians found E-SHAP explanations clearer and more actionable, enhancing decision-making in healthcare.

CONCLUSION

The E-SHAP algorithm provides significant advantages over the regular SHAP, particularly in medicine and finance. E-SHAP is also quicker, which is a 35% improvement for predicting heart disease and a 45% improvement for credit scoring. With this speed, it can now be utilized for real-time decision-making, as required in such domains. E-SHAP produces high-quality explanations. Low values of Mean Absolute Error (MAE) as 0.05 for the heart disease case and 0.07 for the credit score case illustrate how close the outputs of E-SHAP are to those of traditional SHAP. Doctors marked E-SHAP in user experiments as interpretable, obtaining an average rate of 4.5 from 5, while traditional SHAP scored a 3.8. The improved rating indicates that E-SHAP is more interpretable and assistive in decision-making. Overall, E-SHAP improves accuracy as well as model prediction comprehension and is an excellent tool for healthcare and finance analysis. Future improvement could include refining the adaptive sampling techniques further and making E-SHAP available to other domains so that it can be easy to apply and useful.

REFERENCES

[1] To make machine learning a formal and comprehensible discipline, Kim, B., and F. Doshi-Velez (2017). The arXiv preprint ID is 1702.08608.

[2] Singh, S., Ribeiro, M. T., and Guestrin, C. (2016). "Why should I have faith in you?"presents a summary of predictions of each classifier.Proceedings of the 22nd International Conference of ACM SIGKDD: Data Mining and Knowledge Discovery (pp. 1135-1144).

[3] Lee, S.-I. and Lundberg, S. M. (2017). a method to model interpretation of prediction which is reliable. 31st Neural Information Processing Systems Conference Proceedings, NIPS 2017.

[4] Gehrke, J., Sturm, M., Koch, P., Caruana, R., Elhadad, N., and Lou, Y. (2015).

[5] Das, A., Vedantam, R., Parikh, D., Cogswell, M., Selvaraju, R. R., & Batra, D. (2017). Grad-CAM: Gradient-based localization for visual explanations from deep networks.Pages 618–626 of Proceedings of the IEEE International Conference on Computer Vision.

[6] Kundaje, A., Greenside, P., and Shrikumar, A. (2017). Propagating activation differences enables learning of relevant features. 34th International Conference on Machine Learning Proceedings (pp. 3145-3153).

[7] Berrada, M., and Adadi, A. (2018). A survey on explainable artificial intelligence (XAI): A peek into the black box.Access, IEEE, 6, 52138–52160.

[8] Molnar, C. (2019). machine learning that can be understood.The online version is available at http://christophm.github.io/interpretable-ml-book/.

[9] Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., & Inkpen, D. (2018). Improved LSTM for inferring natural language. Arya et al. (2019) investigated salient features of explainability in AI, emphasizing techniques to enhance transparency in computational models.Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, i.e., Volume 1: Long Papers, pages 1657–1668.

[10] Wiegand, T., Samek, W., and Müller, K. R. (2017). Explainable artificial intelligence comprises AI models that are understandable, interpretable, and visualizable. arXiv:1708.08296 preprint.

[11] Arya, V., Hind, M., Hoffman, S. C., Chen, P. Y., Dhurandhar, A., Bellamy, R. K.,.& Zhang, Y.(2019).(2019).A taxonomy of artificial intelligence and a set of explainability techniques.

[12] The former is arXiv preprint 1909.03012.

[13] For explainable machine learning,

[14] Jia, Y., Bhatti, U., Xiang, A., Weller, A., Taly, A., Sharma, S., & Ghosh, J. (2020).

[15] Proceedings of the 2020 Conference on Accountability, Transparency, and Fairness (648-657).

[16] Turini et al. (2018) gave a survey of some of the explanation methods employed in black-box models, pointing out their contribution to the interpretability of complex AI systems. Their paper was published in Surveys of ACM Computing, volume 51, issue 5, covering pages 1–42.

[17] Du, Liu, and Hu (2019) investigated ways to offer machine learning models in an understandable way, prioritizing enhancing their interpretability. Their paper came out in ACM Communications, volume 63, issue 1, pages 68–77.

[18] Knapi, Främling, and Malhi (2019) discussed employing explainable artificial intelligence in medical decision-making systems with special focus on transparency in AI-based healthcare solutions. Their preprint is available on ArXiv with the identifier 2011.14430.