2024 IEEE 20th International Conference on Automation
Science and Engineering (CASE)
August 28 - September 1, 2024. Bari, Italy

Enhancing Robotic Grasp Failure Prediction Using A Pre-hoc Explainability Framework*

Cagla Acun[1], Ali Ashary[2], Dan O. Popa[2] and Olfa Nasraoui[1]

***Abstract*— Enhancing the explainability of machine learning (ML) models is crucial for bridging their use gap for practical applications of robotic autonomy. The common standard for adding explainability has been to use a post-hoc explanation approach. However, post-hoc approaches have recently attracted criticism for their lack of transparency, specifically because the post-hoc methods learn a surrogate model after the predictive model has already been learned, which means that the explanations are not authentically explaining the model's behavior. This study aims to add explainability to the task of robot grasping failure prediction using Factorization Machines as the predictive model and using a novel pre-hoc explainability framework to learn an explainable Factorization Machine model. Unlike post-hoc methods, pre-hoc explainability starts with learning the explainable model before training the black-box model and then provides guidance while learning the latter to make predictions that are faithful to the explanations through a regularization mechanism. Through a detailed case study, we explore the trade-off between prediction accuracy and explanation fidelity and show that our framework is able to make predictions that are more accurate than an explainable white-box model while simultaneously learning a model whose pre-hoc explanations achieve a high level of fidelity relative to the predictions. Results show that our framework can predict the robustness of the grasp with 83% accuracy while explaining that increased effort exerted in Joint 2 of Finger 3 contributes tremendously to producing grasp failure, which is in contrast to increased efforts exerted at Joint 2 of Fingers 1 and 2 and at joint 1 of Finger 3, that all lead to reducing grasp failure.**

I. Introduction

The integration of artificial intelligence into intelligent systems has revolutionized the way industries approach decision-making, automation, and operational efficiency [1]. However, this integration has also raised significant challenges regarding the trustworthiness and interpretability of the underlying algorithms [2], [3]. In particular, autonomous robotic systems, which apply AI and machine learning in uncertain physical systems, often operate as "black boxes" with decision-making processes and failure modes that may not be transparent or easily understood by human operators. This lack of transparency and interpretability hinders the adoption and reliability of intelligent systems in critical applications, where understanding the rationale behind decisions is crucial [4]–[7]. Therefore, providing a clear explanation of such complex models is a significant aspect of increasing trust in machine learning (ML) models of operational success or failure. [8], [9].

Explainable AI techniques have emerged as a promising solution to address the lack of transparency and interpretability in autonomous systems [10] aim to bridge the gap between complex algorithms and human understanding [11]–[13].

Robot grasping is a fundamental task in robotics that involves complex interactions between the robot, its environment, and the object it is grasping. One critical aspect of tasks such as robot grasping is fault diagnostics, which refers to the process of identifying and diagnosing issues or failures within a system. Traditional approaches to fault diagnostics in robot grasping often rely on rule-based systems, physics-based models with incompletely known parameters, or simple sensor thresholding methods, which lack the robustness required for operation in uncertain environments. [14]. ML-based approaches offer the potential to overcome these limitations by using large datasets to learn complex patterns and relationships inherent in the grasping process. By training models on diverse grasping scenarios and corresponding sensor data, ML algorithms can learn to identify subtle deviations indicative of faults or suboptimal grasping strategies [15]. For instance, DexNet is an ML framework under continued development for identifying stable grasp poses from visual information using Convolutional Neural Networks from synthetic or experimental point clouds [16]. Another related approach [17] uses deep reinforcement learning methods in robotic grasping through visio-motor feedback. This approach outperforms baseline methods, enhancing accuracy with a multi-view camera setup.

However, the adoption of ML for fault diagnostics in robot grasping is limited by the lack of interpretability of black-box ML models. In [10], Alvanpour et al. explored the balance between accuracy and interpretability in predicting robot grasp failure by explaining black-box models with post-hoc explanation generation methods, such as Shapley Additive Explanations (SHAP) [18] and LIME [19]. Despite progress towards interpretable fault prediction, post-hoc methods may not be faithful to the original model [20].

To address these challenges, this paper presents a novel approach that leverages a *pre-hoc* explainability framework [21], aiming to enhance the transparency and interpretability of grasp failure prediction. We applied this framework to the analysis of grasp robustness within Shadow's Smart Grasping System [22] using Factorization Machines (FM) [23] as the predictive ML model since FMs capture latent factors of the input variable and their interactions. Unlike our previous

*This work was supported by NSF Grants IIP#1849213 and DRL#2026584. Authors are with the Louisville Automation and Robotics Research Institute (LARRI), University of Louisville, KY 40208, USA.

[1] Cagla Acun and Olfa Nasraoui are with the Knowledge Discovery & Web Mining Lab, Dept. of Computer Science & Engineering, a0acun01,olfa.nasraoui@louisville.edu

[2] Ali Ashary and Dan Popa are with the Next Generation Systems Group, ali.ashary, dan.popa@louisville.edu

post-hoc explainability approach in [10], the proposed pre-hoc method optimizes the predictor model during training to make predictions that are faithful to explanations. Thus, improved fidelity scores can be achieved while maintaining similar levels of accuracy. Results show that our framework can predict the robustness of the grasp with 83% accuracy and display the most influential feature as H1F3J2eff, the effort of the finger 3 joint 2. Our contributions are summarized below:

- We conduct a case study using a novel approach to enhancing the explainability of black-box models, called *pre-hoc explainability*, which leverages the insights provided by an inherently interpretable white-box model to guide the training of the black-box model in a way that preserves its accuracy while enhancing its interpretability.
- Unlike post-hoc explanations, our approach does not rely on input perturbation or post-secondary model learning, thus avoiding the potential pitfalls of surrogate modeling. This makes it more scalable, robust, and reliable in practice.
- We demonstrate the effectiveness and flexibility of our pre-hoc approach on the grasping dataset for the Shadow Hand, showing that a desired prediction accuracy can be attained while ensuring high levels of explanation fidelity.

II. Background

A. Post-hoc Explainability

The current standard explainability approach is Post-hoc explainability, which is widely used to generate explanations for the predictions made by a trained black-box model. However, because the training and explanation generation phases are decoupled (e.g. LIME [19]), they create the risk of having explanations that are a result of some artifacts learned by the model instead of actual knowledge from the data [24]. Post-hoc explanation techniques that rely on input perturbations, such as LIME and SHAP, can, therefore, suffer from unfaithful explanations [20].

Model-agnostic classifiers produce explanations either locally in a few instances or globally across all instances without changing the predictive model itself. One of the most popular algorithms, LIME [19], is an algorithm that approximates a linear regressor or classifier to serve as the explainer. Another common method is SHAP (SHapley Additive exPlanations) [18], which rates each feature according to its contribution to the prediction relative to the contribution of all other input features. By their nature, model-agnostic methods cannot access the internal model state, including model weights or structural details, and their post-hoc explanations are decoupled from the model itself.

B. Pre-hoc Explainability

Unlike Post-hoc methods, Pre-hoc explainability [21] uses pre-trained white-box explanations to guide the learning of a black-box model, as depicted in Figure 1. Table I shows the notation used below. The Explainer function $g \in \mathcal{G}$ serves as a guide to the predictor model f, with the guidance being controlled by minimizing the following distance measure between the explainer and the predictor's outputs, globally:

TABLE I: Variables and Parameters

Symbol	Description
f	Predictor: black-box machine learning model
g	Explainer: white-box machine learning model
p^{ϕ}	Probability distribution of f
p^{θ}	Probability distribution of fg
L_2	Regularization for sparsity
D	Divergence distance measurement
JS	Jensen-Shannon divergence
λ	Explainability regularization coefficient
Z	Grasping dataset

$$\min_{f \in \mathcal{F}} D_{\mathrm{JS}} = \frac{1}{N} \sum_{i=1}^{N} D\left(f\left(x_i\right), g\left(x_i\right)\right), \tag{1}$$

where function D is a divergence distance measurement, specifically the Jensen-Shannon divergence, which is low when the explainer fidelity is high. Finally, the predictive model is learned by minimizing a loss function that combines the loss from the prediction with the Jensen-Shannon divergence:

$$\mathcal{L}_{Pre-hoc} = \mathcal{L}_{BCE} + \lambda_1 D_{\mathrm{JS}} + \lambda_2 \mathcal{L}_2, \tag{2}$$

where $\mathcal{L}_{BCE}$ is the binary cross-entropy loss, λ_1 is an explainability regularization coefficient that controls the trade-off between explainability and accuracy, while λ_2 is a coefficient used for $\mathcal{L}_2$ regularization of model parameters θ to avoid overfitting and exploding gradients. Algorithm 1 shows the steps involved in pre-hoc explainability.

C. Factorization Machines

Factorization machines (FMs) [23] are supervised learning models that can be applied to a wide range of prediction tasks while reliably estimating model parameters under large quantities of sparse data, enabling the model to be trained with very few data points. The model equation for a factorization machine of degree d = 2 is defined as:

$$\hat{y}(x) = \underbrace{w_0}_{\text{Term 1}} + \underbrace{\sum_{i=1}^{n} w_i x_i}_{\text{Term 2}} + \underbrace{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j}_{\text{Term 3}} \tag{3}$$

where the model parameters that have to be estimated are:

$$w_0 \in \mathbb{R}, w_i \in \mathbb{R}^n, v_i \in \mathbb{R}^{n \times k}, \tag{4}$$

And $\langle \cdot, \cdot \rangle$ is the dot product of two vectors of size k:

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \sum_{f=1}^{k} v_{i,f} v_{j,f} \tag{5}$$

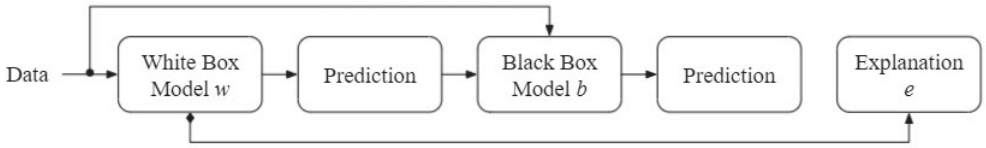


Fig. 1: Training Phase of Pre-hoc Explainability Framework

A row v_i within V represents the i^{th} variable with k factors, where k is a hyperparameter that defines the dimensionality of the factorization.

D. Transparency in Factorization Machines

One of the key drawbacks of Factorization Machines (FMs) is it's lack of interpretability. In FMs, the variable interactions are modeled by a polynomial expansion, which is difficult to explain. In 3, Term 1 and Term 2 are classical regression terms that are explainable. However, Term 3 (v_i, v_j) adds opacity, thus reducing the transparency of the predictions.

Algorithm 1 Pre-hoc Explainability Framework

Require: Black-box model f_θ, white-box model g_ϕ, input instance x, true label y, and parameter λ_1 for weighting the divergence term.

procedure PREHOCEXPLAINABILITY(f_θ, g_ϕ, x, y, λ_1)

for each x_i in X_{train} **do**

Compute $p^\phi = g_\phi(x_i)$ ▷ Predictions from white-box model

Compute $p^\theta = f_\theta(x_i)$ ▷ Predictions from black-box model

Compute $L_{JS} = \text{JS}(p^\theta, p^\phi)$ ▷ Using JS divergence

Compute $L_{BCE} = \text{BinaryCrossEntropy}(p^\theta, \text{y})$

$L_{total} = L_{BCE} + \lambda_1 \cdot L_{GJS}$

Update f_θ using gradient descent: $\theta \leftarrow \theta - \alpha \nabla_\theta L_{total}$

end for

end procedure

III. Problem Formulation

We focus on a robot's hand with three fingers, including information about the joints' position, velocity, effort (torque) of each finger, and stability of the grasp for an object. Our aim is to predict grasp failure from the position, velocity, and effort measurements of each of the three joints in each of the three fingers. These measurements are collected into features that are named after the combination of hand (only Hand 1 is used), finger, join, and either position, velocity, or effort, as summarized in the following nomenclature.

- $H1$: Hand one, indicating the only hand used in the simulation.
- $F1, F2, F3$: Fingers on the hand, where each finger has three joints.
- $J1, J2, J3$: Joints in each finger, with each joint having measurements for position (pos), velocity (vel), and effort ($effort$).

Hence $H1FjJk$ indicates joint k of Finger j of Hand 1. The dataset for a single experiment e_i can be represented as a matrix:

$$M_{e_i} = \begin{bmatrix} pos_{H1F1J1} & vel_{H1F1J1} & effort_{H1F1J1} \\ \vdots & \vdots & \vdots \\ pos_{H1F3J3} & vel_{H1F3J3} & effort_{H1F3J3} \end{bmatrix}$$

where pos_{H1FjJk}, vel_{H1FjJk}, and $effort_{H1FjJk}$ represent the position, velocity, and effort measurements of joint k in finger j of hand one, respectively.

The grasp robustness R for each experiment is computed based on the variation of the distance between the palm and the ball during the shake, as shown in Figure 2, denoted as:

$$R(e_i) = f(\Delta dist_{palm-ball}(e_i))$$

Where f is a function that computes the robustness based on the distance variation $\Delta dist_{palm-ball}$ during experiment e_i. By having all these features as a dataset $\mathcal{Z}$, then, let $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{\mathcal{N}} \subset \mathcal{Z}$ be a sample from a distribution $\mathcal{D}$ in a domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the instance and $\mathcal{Y}$ is the label set. We learn a differentiable predictive function $f \in \mathcal{F} : \mathcal{X} \to \mathcal{Y}$ together with a transparent function $g \in \mathcal{G} : \mathcal{X} \to \mathcal{Y}$ defined over a functional class $\mathcal{G}$. We refer to functions f and g as the predictor and the explainer, respectively, throughout the paper. $\mathcal{G}$ is strictly constrained to be an inherently explainable functional set, such as a set of linear functions or decision trees. We assume that we have a distance function $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ such that $d(y, \hat{y}) = 0 \longleftrightarrow y = \hat{y}$, which measures the point-wise similarity between two probability distributions in $\mathcal{Y}$ and can be used to optimize f and g.

Instead of learning a post hoc white-box model, we learn a model that is explainable from the start and then let this explainer model guide the predictor. We use the Pre-Hoc Explainable Predictive Framework, where the white box model regularizes the black box model for higher fidelity (Figure 1).

IV. Experiments

We evaluate the performance of our approach on the grasping dataset obtained from Shadow's Smart Grasping System [22] simulation with ROS [25] and Gazebo [26] environment using the Smart Grasping Sandbox [22], depicted in Figure 2, and containing three 3-DOF fingers. The dataset has been annotated with an objective grasp of consistency and contains

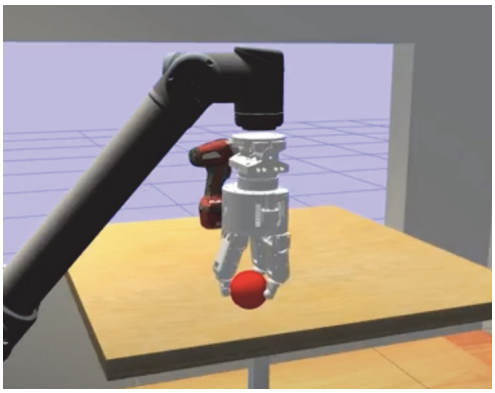


Fig. 2: Shadow Robot in the Smart Grasping Simulation

various data obtained from the joints (position, velocity, and effort), containing about 54,000 unique data points and 29 measurements for each experiment. The classification target is the predicted grasp robustness. Moreover, the output is discretized to 1 for a stable grasp and 0 for an unstable grasp. A grasp is considered stable if the robustness value exceeds 100.

A. Data Preprocessing

The discretization process can be defined as a function D that maps each robustness value r to a discrete label l, as follows:

$$D(r) = \begin{cases} 0 & \text{if } r < 100 \\ 1 & \text{if } r \geq 100 \end{cases}$$

The dataset Z is normalized to standardize the feature values, ensuring that each feature contributes equally to the analysis.

Therefore, for every element r in Z, we apply D to obtain a binary label indicating whether the grasp robustness is below or above a threshold of 100, effectively categorizing the data into two classes: less robust (0) and more robust (1). This process simplifies the target for predictive modeling, focusing on the binary classification of grasp robustness.

B. Experimental Setup and Evaluation

The dataset was randomly divided into training, validation, and test sets with an 80:10:10 ratio. Model performance and fidelity were assessed using AUC on the validation and test datasets after training with a learning rate of 0.001 and L_2 regularization until validation accuracy stabilized for at least ten epochs. We used the **Area under the ROC Curve** $AUC(f_\theta, \hat{y})$ as our primary metric for classification accuracy, and **Fidelity** as our transparency metric. Fidelity reflects the descriptive accuracy of our explanation method against the black-box classifier [27]. It is measured by computing the AUC between the original black-box predictions and the explainer model's predictions $AUC(f_\theta, g_\phi)$, with the explainer model.

Our approach was benchmarked against a black-box (BB) version using Factorization Machines, while the white-box (WB) model was implemented as a sparse logistic regression model for its inherent explainability. We adopted PyTorch for implementation, leveraging Adam [28] for optimization and binary cross-entropy for loss calculation. The models were trained with a batch size of 2056, and λ_1 was varied across 0.05, 0.01, 0.1, 0.25, 0.5, 1 to determine the optimal regularization weight via a validation set. The final evaluation was conducted by retraining the models with their optimal configurations and assessing them on the test set.

V. Results

A. Ablation Study

We did an ablation study to compare the regularized black-box (Pre-hoc Predictor) and the non-regularized black-box (BB Predictor) in Table II. The explainer model used was a white-box (WB) model, while the predictor model was a black-box (BB) model. In Table II, the BB Predictor and Pre-hoc Predictor models showed similar accuracy scores with almost no differences in AUC. Also they maintained a higher AUC score compared to the WB Explainer model, indicating better prediction accuracy. In terms of fidelity, which measures the explainer's accuracy in mimicking the black-box model's decisions, while maintaining the same accuracy score Pre-hoc Predictor showed 4.3% increase on fidelity. The Pre-hoc BB over regularized model achieved the highest score 0.9560 which is approximately 27% increase, significantly outperforming other models, see Figure 3 (b). This suggests that while the Pre-hoc BB model has the best predictive and explainer performance, the Pre-hoc BB over-regularized model offers the most reliable explanations, highlighting a trade-off between prediction accuracy and explainability fidelity.

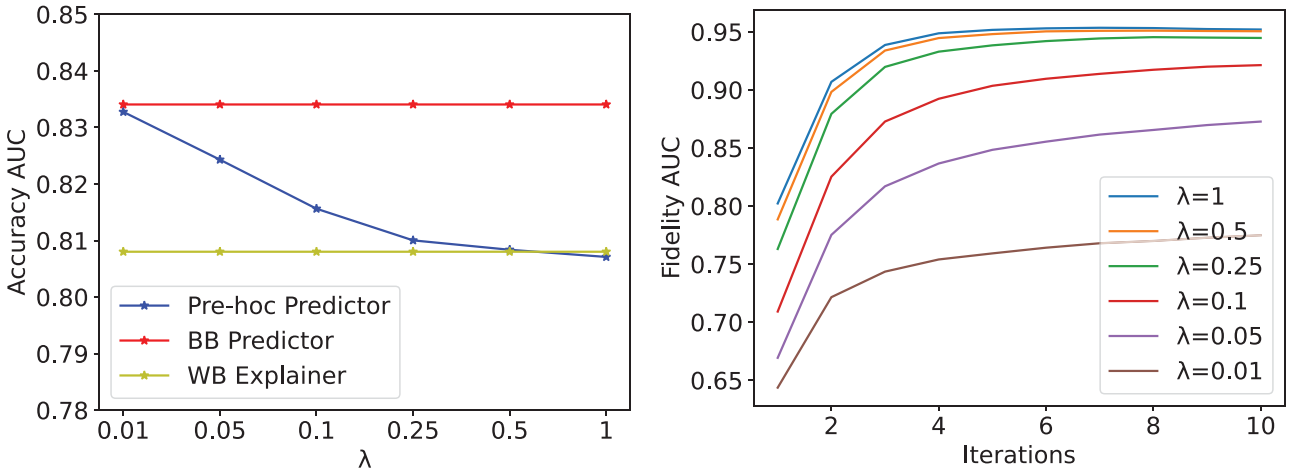
TABLE II: Model comparison in terms of prediction accuracy and fidelity of explainability. The explainer is the white-box model, BB is the predictor black-box model. The best results are in **bold**. Higher AUC and Fidelity is better.

Dataset	Grasping	
Model	AUC ↑	Fidelity ↑
Explainer WB	0.8080	-
Predictor BB	**0.8340**	0.7532
Pre-hoc BB λ=0.01	0.8327	0.7860
Pre-hoc BB λ=0.05	0.8242	0.8718
Pre-hoc BB λ=0.1	0.8120	**0.9560**

B. Transparency and Accuracy Trade-off

The experiments collectively highlight the inherent challenge in optimizing both model accuracy and transparency, suggesting that enhancing one aspect can often lead to compromises in the other. This trade-off is critical in developing and evaluating intelligent systems, where both accuracy and explainability are essential for trust and reliability.

Specifically, Figure 3 shows that from $\lambda = 0.1$, the AUC decreased 0.834, for $\lambda = 1$ to 0.812, but fidelity improved



(a) (b)

Fig. 3: Experimental Results in Accuracy AUC all models (a) and Fidelity AUC for Pre-hoc Predictor (b) for different λ's. Pre-hoc Predictor is the regularized predictor model, BB is the baseline black-box predictor model, WB is the explainer model.

from 0.753 to 0.956, which is approximately a 27% increase, showcasing a substantial gain in model transparency at a modest cost to accuracy. Thus, the ability to adjust the λ value accuracy and transparency demonstrates our framework's flexibility in balancing the tradeoff. In fact, we obtain a good balance at λ=0.05, where the AUC accuracy and fidelity achieve favorable scores of 0.82 and 0.87, respectively.

C. Global Explainability

Global explainability addresses the need to understand a machine learning model's decision-making on a broad level, containing the entire model rather than individual predictions. The top 10 features from the Pre-hoc framework, as shown in Figure 4, indicate the relative importance of each feature and its contribution to predicting grasp robustness. Figure 4 shows that the most influential feature is $H_1F_3J_2eff$, joint 2's effort in finger 3, which increases the model score by 0.5. It shows a strong positive correlation, suggesting a significant impact on grasp stability. In contrast, $H_1F_2J_2eff$ joint 2 effort in finger 2 demonstrates a negative influence on robustness by decreasing the model score by 0.25. Also, effort (e.g., torque) consistently has more effect on grasping results than the grasping velocity. This disparity in feature impact highlights the complex interaction between joint effort and velocity in determining the successful execution of a grasp. The analysis of these top features not only provides insights into the decision of the grasping process but also reinforces the value of explainable AI in enhancing our understanding.

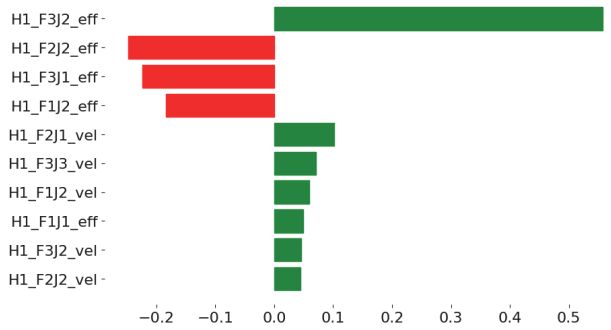


Fig. 4: Top 10 Feature Importance: Global explanation from the Pre-hoc framework, showing that increased effort exerted in Joint 2 of Finger 3 contributes tremendously to producing grasp failure, which is in contrast to increased efforts exerted at joint 2 of Fingers 1 and 2 and at joint 1 of Finger 3, that lead to reducing grasp failure.

VI. Conclusion

We demonstrated the integration of a pre-hoc explainability framework using Factorization Machines for enhancing the explainability of machine learning models in autonomous robotic grasping systems. Through experiments and analysis, we have illustrated the inherent trade-off between accuracy and transparency, revealing that increased regularization can significantly enhance model fidelity without substantially compromising predictive accuracy. This research contributes valuable insights into the optimization of autonomous decision-making processes, offering a pathway to more interpretable and user-trustworthy intelligent systems.

The current work has several limitations. Common explainable AI techniques often prioritize model-internal explanations, overlooking the valuable insights that domain experts or end-users can provide. So, one of the main limitations is that explanations are only model-internal. Another limitation is that only feature importance scores are used, which is the most commonly used explanation format. Future

work should explore methods for integrating domain knowledge and feedback into the explanation generation process, ensuring that the explanations are meaningful relative to domain-specific requirements and user expectations.

REFERENCES

[1] J. P. Nelson, J. B. Biddle, and P. Shapira, "Applications and societal implications of artificial intelligence in manufacturing: A systematic review," *arXiv*, 07 2023.

[2] S. Ornes, "Peering inside the black box of ai — proceedings of the national academy of sciences," 05 2023.

[3] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 03 2021.

[4] W. Samek, T. Wiegand, and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 08 2017.

[5] S. Sengupta, S. Basak, P. Saikia, S. Paul, V. Tsalavoutis, F. D. Atiah, V. Ravi, and A. Peters, "A review of deep learning with special emphasis on architectures, applications and recent trends," *Knowledge-Based Systems*, vol. 194, pp. 105 596–105 596, 04 2020.

[6] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, "Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond," *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3197–3234, 09 2022. [Online]. Available: https://arxiv.org/abs/2103.10689

[7] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, no. none, 01 2022. [Online]. Available: https://arxiv.org/pdf/2103.11251.pdf

[8] A. Bennetot, I. Donadello, A. E. Qadi, M. Dragoni, T. Frossard, B. Wagner, A. Saranti, S. Tulli, M. Trocan, R. Chatila, A. Holzinger, A. d'Avila Garcez, and N. Rodruedguez, "A practical guide on explainable ai techniques applied on biomedical use case applications," 01 2022.

[9] J. Zhang and R. X. Gao, "Deep learning-driven data curation and model interpretation for smart manufacturing," *Chinese journal of mechanical engineering*, vol. 34, no. 1, 07 2021.

[10] A. Alvanpour, S. K. Das, C. K. Robinson, O. Nasraoui, and D. Popa, "Robot failure mode prediction with explainable machine learning," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, 2020, pp. 61–66.

[11] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Şensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: A survey of results," 08 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8397411/

[12] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," *arXiv*, 05 2018. [Online]. Available: https://arxiv.org/abs/1806.00069v2

[13] W. Samek, "Explainable artificial intelligence," 02 2023. [Online]. Available: https://www.itu.int/en/journal/001/Pages/05.aspx

[14] C. Wang, X. Zhang, X. Zang, Y. Liu, G. Ding, W. Yin, and J. Zhao, "Feature sensing and robotic grasping of objects with uncertain information: A review," *Sensors*, vol. 20, no. 13, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/13/3707

[15] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler, "Deep learning algorithms for bearing fault diagnostics—a comprehensive review," *IEEE Access*, vol. 8, pp. 29 857–29 881, 2020.

[16] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.

[17] S. Joshi, S. Kumra, and F. Sahin, "Robotic grasping using deep reinforcement learning," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, 2020, pp. 1461–1466.

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[19] 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, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[20] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 180–186.

[21] C. Acun and O. Nasraoui, "In-training explainability frameworks: A method to make black-box machine learning models more explainable," in *2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2023, pp. 230–237.

[22] Shadow Robot, "Shadow's Smart Grasping System," 2017. [Online]. Available: https://github.com/shadow-robot/smart_grasping_sandbox

[23] S. Rendle, "Factorization machines," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, ser. ICDM '10. USA: IEEE Computer Society, 2010, p. 995–1000. [Online]. Available: https://doi.org/10.1109/ICDM.2010.127

[24] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The dangers of post-hoc interpretability: Unjustified counterfactual explanations," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 7 2019, pp. 2801–2807.

[25] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system." [Online]. Available: https://www.ros.org

[26] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, vol. 3, 2004, pp. 2149–2154 vol.3.

[27] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, oct 2019.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.