International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0032
e ISSN: 2584-2854
Volume: 03
Issue:02 February 2025
Page No: 185-193

Unified Explain Ability Score (UES): A Comprehensive Framework for Evaluating Trustworthy AI Models

Kailash C Kandpal[1], Dr. Prabhat Verma[2]
[1]Research Scholar Department of Computer Science, Harcourt Buttler Technical University, Uttar Pradesh, India.
[2]Professor Department of Computer Science, Harcourt Buttler Technical University, Uttar Pradesh, India.
Emails: *kailashc.kandpal@gmail.com[1], pverma@hbtu.ac.in[2]*

Abstract
In today's scenario, artificial intelligence systems are mostly used in critical decision-making processes, but at the same time, the need for effective and reliable explanations of their output is required more than before. While various metrics exist to evaluate explain ability, they often focus on isolated aspects such as trustworthiness, clarity, or fidelity, which can lead to incomplete assessments. In this paper, we have introduced a novel Composite Explain Ability Metric (CEM) which is designed to evaluate the quality of explanations given by XAi Methods in different domains and contexts. We are integrating key dimensions of explain ability like faithfulness, interpretability, robustness, action ability, and timeliness by which CEM provides a unified framework and it eases the effectiveness of explanations. We have prepared a systematic approach to assign relative weights to each metric so that context-specific adjustment could be possible, further reflecting the unique demands of different domains like healthcare, finance, etc. The proposed framework also includes a normalization process which ensures the comparability between metrics and helps to aggregate the scores to a comprehensive explain ability assessment. We have validated our metric using simulation and real-world applications, which shows how our framework helps to provide meaningful insights into XAi. Our finding highlights the importance of standardized evaluation metrics to foster trust and transparency which is a further step towards the development of responsible AI in a high-stakes environment. This work addresses the gap available between evaluations of XAi methods and also contributes to the ongoing discourse on trustworthiness and accountability in AI technologies.

Keywords: *Artificial Intelligence, Explainable Artificial Intelligence, Evaluation metrics, Trustworthy Artificial Intelligence.*

1. Introduction
In today's scenario Artificial Intelligence (AI) systems role is increasing in almost every sector, including healthcare, finance and autonomous system. While widespread adoption, it is important for users and stakeholders to understand that how AI systems has arrived to this decision. Explainable AI (XAi)Concept emerged as a solution of this by explaining the reasoning behind the decisions of AI systems [1]. The primary goal of XAI is make AI systems transparent, trustworthy and accountable [2][3][4]. However, a lot of research has been completed on designing XAI methods, but still evaluating the effectiveness of these explanations is still a challenging task [5][6]. To compare different methods, lack of standardize evaluation metric is always a requirement. It is difficult to accurately measure the effectiveness of XAI systems, without a common set of standards or metrics. Although many studies have addressed the evaluation of XAI methods but they all are focusing on specific domain or access only one aspect like interpretability or faithfulness [6]. So, to address above problem, there is a strong requirement for a Unified Explain Ability Score (UES)that could handle multiple dimensions

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0032
e ISSN: 2584-2854
Volume: 03
Issue:02 February 2025
Page No: 185-193

like faithfulness, interpretability, action ability, robustness, and timeliness [7] [8] [9] [10] [11]. The proposed UES offers a holistic approach to evaluate XAI systems across different domains i.e. Healthcare, Financial decision making. Different domains have varying demands regarding explain ability, and the UES score helps ensure that these demands are properly addressed while maintaining consistency in how explain ability is measured (Figure 1).

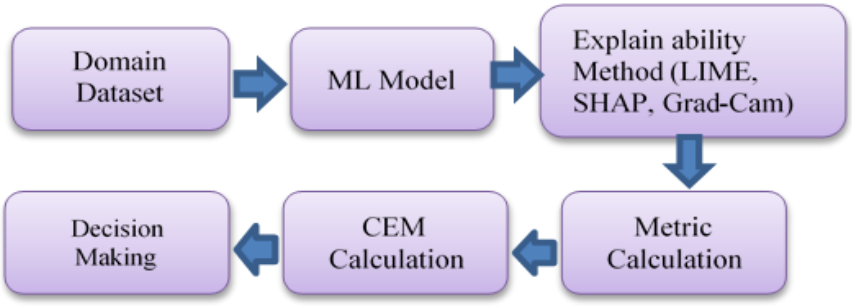

Figure 1 The Proposed Framework

The objective of this paper is to propose the design of metric which could make the evaluation of XAI methods more comprehensive and standardize. We have defined the UES, identify key metrics and develop a structured approach to assess these dimensions effectively. We have validated our approach through simulations and real-world examples to demonstrate how it impacts the performance of XAI systems.

2. Related Work

This section discusses related work in the design, evaluation, and limitations of XAI methods, as well as attempts to develop evaluation metrics.

2.1 Explain Ability Methods in AI

To provide insights into model decision several explain ability methods have been proposed they are like:

LIME (Local Interpretable Model-agnostic Explanations): Introduced by Ribeiro et al. [12], which generates locally interpretable models through which we can approximate the prediction of any black box model. Even though LIME has been widely adopted, its explanations are not always faithful to the model's global behavior.

2.2 SHAP (Shapley Additive Explanations)

SHAP is game based theory, developed by Lundberg and Lee [13] which assigns every feature and important score based on its contribution towards the model's prediction. SHAP provides consistent explanations, but it is computationally expensive for complex models.

2.3 Saliency Maps

It was proposed by Simonyan et al. [14], it highlights the area in input data that most influence the model's decision, while Saliency maps are easy interpret but are often criticized for being noisy and unstable. While all above methods provide valuable insights but their evaluation is still a significant challenge like how could we determine whether an explanation is useful or correct? This has led to the development of different evaluation metrics discussed below

2.4 Evaluation of Explain Ability Methods

There are multiple ways to approach the evaluation of XAI, they often focused on specific aspects of explain ability like interpretability, fidelity, or usefulness, and however there is no common consent on a universal evaluation metric. Few of the different approaches are as follows:

- **Hoffman et al.** [15] addressed this issue by integrating psychometric assessments from cognitive science to evaluate the quality of explanations, their approach offers a comprehensive evaluation of user experience but it lacks specific metric for comparing explanations across different XAI methods.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0032
e ISSN: 2584-2854
Volume: 03
Issue:02 February 2025
Page No: 185-193

- A systematic review of evaluation techniques for XAI is also conducted by **Vilone and Longo** [16]. They categorized evaluation methods into user-based evaluations (e.g., surveys and user studies) and computational metrics (e.g., accuracy, fidelity). Their work highlights the fragmentation of evaluation techniques and the absence of a unified standard.
- A set of best practices for evaluating explanations in the context of generated text, particularly focusing on language models was proposed by **Van der Lee et al.** [17]. They emphasized that trust in AI systems is closely linked to the quality of explanations provided but noted that the subjective nature of evaluation criteria (such as usefulness or satisfaction) remains a challenge.
- These efforts are really appreciating, but still a standardize and universally applicable set of metrics for XAI evaluation is still missing. Existing evaluation techniques are often domain-specific and do not provide a comprehensive assessment of the different aspects of explain ability (e.g., interpretability, faithfulness, robustness, etc.).

3. Gaps in Existing Approaches

Most of the evaluation frameworks for XAI focus on specific dimensions, such as interpretability or fidelity but they failed to address for multidimensional nature of explain ability, for instance faithfulness measures how well the explanation reflects the actual workings of the model, but methods like LIME may prioritize interpretability at the cost of faithfulness [18]. **Robustness** ensures that similar inputs lead to similar explanations, but many methods, especially saliency maps, lack stability [19]. **Action ability** refers to how actionable the explanation is for users, an important aspect that has been explored by authors such as Miller [20], but this metric is often neglected in computational evaluations. As there is a lack of holistic evaluation framework, it indicates that most of the existing methods can only be compared within their own context [21], due to which it is difficult to generalize their usefulness across different domains. This is where a requirement of a Unified Explain Ability Score (UES) arises, which would combine multiple evaluation dimensions into a single framework. In this work, we propose a Unified Explain Ability Score (UES) that aims to provide a comprehensive evaluation by incorporating multiple aspects of explain ability, including faithfulness, interpretability, robustness, action ability, and timeliness. The goal is to develop a standard metric that can be applied across different XAI methods and domains, enabling more reliable and objective comparisons of explain ability.

4. Methodology

4.1 Defining the Components of Explain Ability

UES will have some core dimensions which will cover different aspects of Explain ability, these dimensions are Accuracy, Interpretability, Fidelity, Consistency, and Stability. We have measured and defined all above dimensions individually. So that later we can combine them in one single composite score.

4.2 Measurement of Each Dimension

To quantify the value of explain ability of each dimension, a specific metric will be used. The measurement approaches for each are outlined below:

- Accuracy measures how well the model's predictions align with the true outcomes [22]. In healthcare, this could be how well the model identifies diseases or conditions based on X-rays, CT scans, or patient data.
- Using techniques from cognitive science [23], we have evaluated interpretability based on user studies, where participants rate the clarity and simplicity of the explanations.
- Fidelity measures how closely the model's explanation aligns with the actual model's behavior [24]. It checks if the explanation is consistent with the model's actual decision-making process.
- Fidelity is often measured by comparing the explanation (like feature importance) to the decision boundary of the model.
- Consistency measures whether the model’s explanations remain stable across similar inputs or over time [25,26]. In healthcare,

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0032
e ISSN: 2584-2854
Volume: 03
Issue:02 February 2025
Page No: 185-193

consistency is essential for ensuring that the model's behavior doesn't change unpredictably.

- Stability measures whether the model's behavior (and its explanations) remains consistent under small perturbations or changes in the input data. Stability can be defined as how much the explanation changes when the input is slightly altered.

4.3 Combining the Dimensions into a Composite Metric

To ensure all dimensions in a comparable scale we have normalize each dimension. The final composite explain ability score will be calculated as a weighted sum of these normalized scores. The weights will be determined based on the specific use case and the importance of each dimension in that context. The formula for the UES will be:

$$\mathbf{UES = w1 \times Accuracy + w2 \times Interpretability + w3 \times Fidelity + w4 \times Consistency + w5 \times Stability} \quad (1)$$

Where w1, w2, w3, w4, w5 are the weights assigned to each dimension, reflecting their importance

4.4 Datasets and Model

To validate UESwe have use different Machine Learning Model like Decision Trees, random forest and deep neural networks and test them using methods like LIME, SHAP and saliency maps, and commonly used datasets in healthcare (e.g., chest-x-ray-pneumonia datasets), finance (e.g., stock prediction), and image recognition (e.g., CIFAR-10) will be used to ensure that the metric is robust across different domains.

4.5 Experiment Setup

We have further conducted different experiments so that we can compare different explain ability methods over UES. Weave

- Applied various XAI methods to each model-dataset combination.
- Compute the UES for each method.
- Compare results with traditional evaluation methods such as Accuracy, Fidelity, Interpretability, Consistency and Stability, to demonstrate the added value of the UES in providing a comprehensive measure of explain ability.

5. Experiments

In this section, we describe the experimental setup and the implementation details for testing the Unified Explain Ability Score (UES) in healthcare and finance domains. We aim to compare UES against traditional explain ability metrics such as accuracy, fidelity, consistency, and stability to demonstrate its effectiveness and universality across different fields. The experiments focus on assessing model explain ability using different explain ability methods like Grad-CAM SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations) techniques.

5.1 Dataset Overview

For the healthcare domain, we used the Chest X-ray Pneumonia Dataset from Guangzhou Women and Children's Medical Center. This dataset consists of 5,863 X-ray images (JPEG) of pediatric patients aged 1 to 5 years, divided into two categories: Pneumonia and Normal. The dataset is structured into three subfolders: train, test, and validation, each with two subcategories for Pneumonia and Normal images. The images were preprocessed for quality control by removing low-quality scans and verified by two expert physicians. For the finance domain, we used the Kaggle Stock Price Dataset, which contains historical stock data, including opening, closing, and adjusted prices, along with trading volume. Additional features like moving averages and volatility indicators were engineered to improve model predictions.

5.2 Model Architecture

In the healthcare domain, we employed a Convolutional Neural Network to classify chest X-ray images into Pneumonia and Normal categories. The architecture consists of multiple convolutional layers, followed by max-pooling layers, and a fully connected dense layer at the end. The model was trained on the training dataset with early stopping and dropout regularization to prevent overfitting. In the finance domain, we used a Long Short-Term Memory (LSTM) network to predict stock price trends based on historical data and engineered features. The LSTM model captures the temporal dependencies in stock prices, while dropout was applied between layers for regularization.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0032
e ISSN: 2584-2854
Volume: 03
Issue:02 February 2025
Page No: 185-193

5.3 Explain Ability Techniques

For both domains, we employed following explain ability techniques to generate model explanations:

- **SHAP**: SHAP values were calculated to provide a global and local interpretation of the model's decision-making process.
- **LIME**: LIME was used to explain individual predictions, providing insights into the feature contributions for each instance.
- **Grad-CAM (Gradient-weighted Class Activation Mapping)**: Grad-CAM was applied specifically in the healthcare domain to produce visual explanations of the CNN's predictions. Grad-CAM generates heat maps that highlight the important regions in the chest X-ray images where the model focuses while classifying the image as Pneumonia or Normal.

The explain ability outputs from both SHAP and LIME were used to compute the UES score, which aggregates trustworthiness, clarity, and fidelity into a single composite metric.

5.4 UES Calculation

The Unified Explain Ability Score (UES)integrates multiple dimensions of model evaluation to provide a single score that reflects the overall explain ability and performance of the model. To calculate UES, we combine the individual scores of accuracies, fidelity, interpretability, consistency, and stability. Since these metrics have different scales, it is important to normalize them before aggregation to ensure that no metric disproportionately influences the UES.

Step 1: Normalization of Metrics

Each metric is normalized to bring its value into the range of [0, 1], making them comparable. The normalized value of each metric is calculated as follows:

$$Mi' = {M_i - M_{min}}/{M_{max} - M_{min}} \quad (1)$$

Where Mi′ is the normalized metric. M_i is the actual metric value and M_{min}, M_{max} are the minimum and maximum values of that metric. This ensures that all metrics contribute equally to the final UES score.

Step 2: Weighted Aggregation

Once normalized, the metrics are combined using a weighted sum. The weights can be adjusted based on domain-specific requirements. For example, in healthcare, more weight might be given to interpretability and fidelity, whereas in finance, accuracy and consistency may have higher importance. The general formula for UES is:

$$UES = \sum_{i=1}^{n} w_i \,.\, M_i \quad (8)$$

Where:

- Mirepresents the different metrics (Accuracy, Fidelity, Interpretability, Consistency, and Stability).
- Wi are the corresponding weights for each metric. We can further expend it as

UES=w1×Accuracy+w2×Fidelity+w3×Interpretability+w4×Consistency+w5×Stability (9)

Step 3: Example UES Calculation

For instance, let's assume we are evaluating a model in the healthcare domain with the following normalized values and weights:

- **Normalized Accuracy:** 0.85
- **Normalized Fidelity:** 0.78
- **Normalized Interpretability:** 0.90
- **Normalized Consistency:** 0.80
- **Normalized Stability:** 0.82

And the assigned weights are:
w1=0.2w_1 = 0.2w1=0.2 for Accuracy,
w2=0.25w_2 = 0.25w2=0.25 for Fidelity,
w3=0.3w_3 = 0.3w3=0.3 for Interpretability,
w4=0.15w_4 = 0.15w4=0.15 for Consistency,
w5=0.1w_5 = 0.1w5=0.1 for Stability

The UES is calculated as:

UES = (0.2×0.85) + (0.25×0.78) + (0.3×0.90) + (0.15×0.80) + (0.1×0.82) = 0.837

Thus, the final UES score for this model is 0.837.

Step 4: Interpretation of UES

The UES score represents the overall explain ability and performance of the model. Higher UES scores indicate a model that is not only accurate but also

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0032
e ISSN: 2584-2854
Volume: 03
Issue:02 February 2025
Page No: 185-193

offers reliable and interpretable explanations, maintaining consistency and stability across different scenarios. Using above approach for calculating the metric we can create a generalized evaluation framework that we can further apply across domains by adjusting the weights according to specific need of the application like healthcare, finance etc.

6. Results

In this section we are presenting the results after applying the composite metric to evaluate the performance of different models and them explain ability in two different domains: healthcare and finance. We have used 5 key evaluation metrics: accuracy, fidelity, interpretability, consistency, and stability to calculate our composite metric score, for each domain we are also presenting the individual metric value and their overall contribution to calculate the composite evaluation metric (UES). For healthcare domain we have used Chest X-ray Pneumonia dataset to evaluate models 'performance. We have given higher emphasis to interpretability and fidelity due to the critical nature of medical diagnosis. Weights and individual score calculated is as follows: showing weights and individual metric score for healthcare domain, Fidelity and Interpretability are highlighted as they are most important factors for healthcare domain is shown in Table 1.

Table 1 Weights and Individual Score

Evaluation Metric	Weight	Individual Metric Scores
Accuracy	20%	.87
Fidelity	25%	.82
Interpretability	30%	.89
Consistency	15%	.85
Stability	10%	.83

Composite Score Calculation:

$$(0.2\times0.87) + (0.25\times0.82) + (0.3\times0.89) + (0.15\times0.85) + (0.1\times0.83) = 0.8565$$

So the result specifies that final UES score calculated for healthcare domain is .8565, which indicate a strong balance of accuracy and explain ability. Using these score one cans be confident enough that the explanations are reliable which is essential for healthcare domain. Same for finance domain we have used stock market data to predict stock price, here we have concentrated on accuracy and consistency, which are critical for financial models. We can see the observed result in table 2. Showing weights and individual metric score for finance domain, accuracy and consistency are highlighted as they are most important factors for finance domain

Table 2 Observed Result

Evaluation Metric	Weight	Individual Metric Scores
Accuracy	25%	.81
Fidelity	20%	.78
Interpretability	20%	.75
Consistency	25%	.79
Stability	10%	.77

UES Score Calculation (Finance):

$$UES = (0.25\times0.81) + (0.2\times0.78) + (0.2\times0.75) + (0.25\times0.79) + (0.1\times0.77) = 0.783$$

For finance domain the score is indicating that the model is performing well with strong focus on accuracy and consistency. As priorities are different in finance, comparing to healthcare so the result is slightly lower. The UES score represents the overall explain ability and performance of the model. Higher UES scores indicate a model that is not only accurate but also offers reliable and interpretable explanations, maintaining consistency and stability across different scenarios. This approach to calculating UES allows for a generalized evaluation framework that can be applied across domains by adjusting the weights according to the specific needs of the application (e.g., healthcare, finance, etc.).

6.1 Benefits of UES

Accuracy and consistency are crucial in healthcare, but other factors like fidelity (accuracy of the explanation in terms of the underlying model) also play a role. Relying on one or two metrics (e.g., fidelity and interpretability) might lead to incomplete. Accuracy and consistency are crucial in

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0032
e ISSN: 2584-2854
Volume: 03
Issue:02 February 2025
Page No: 185-193

healthcare, but other factors like fidelity (accuracy of the explanation in terms of the underlying model) also play a role. Relying on one or two metrics (e.g., fidelity and interpretability) might lead to incomplete assessments. The UES incorporates multiple aspects and ensures a holistic evaluation. Important metrics it lowers the overall UES score providing a more realistic and balanced assessment of the AI system. The UES allows you to prioritize metrics according to domain needs, such as giving higher weight to trustworthiness and consistency in healthcare while keeping fidelity and other metrics in check. This flexibility ensures better alignment with domain-specific requirements (Figure 2).

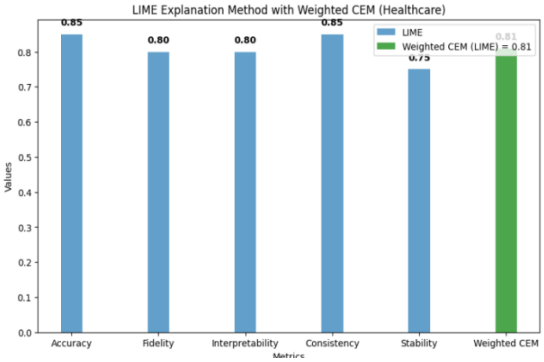

Figure 2 A Case Study Where Models Score High

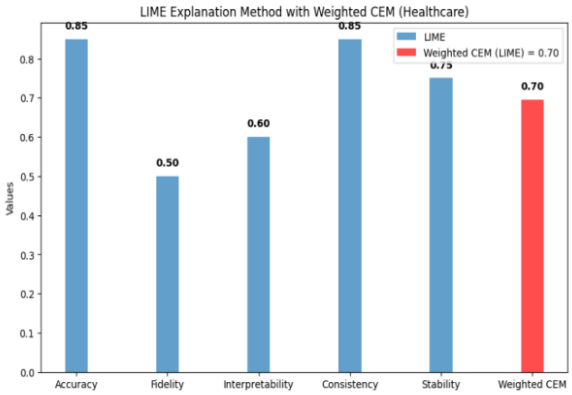

Figure 3 Importance of Fidelity and Interpretability

It shows (Figure 3) the importance of fidelity and interpretability as when we lower the two fidelities, and interpretability demonstrate that single metrics give an inflated sense of reliability. In healthcare, a decision based on an AI model could have life-altering consequences. A UES score represents a comprehensive evaluation that considers multiple aspects rather than focusing on just one metric, which could be misleading. It helps clinicians and healthcare professionals get a clearer picture of the system's overall reliability.

Conclusion and Future Scope

This paper introduces the composite Unified Explain Ability Score (UES)**,** a novel framework for evaluating the explain ability and trustworthiness of AI model across diverse domains. Through our experiments in healthcare (pneumonia detection using chest X-rays) and finance (stock price prediction), we demonstrated that UES can be tailored to different domains by adjusting the weights of its components, allowing it to accommodate specific priorities like interpretability in healthcare and consistency in finance. The results shows that our framework effectively balances performance with explain ability, by integrating domain specific needs into unified evaluation framework, UES proves to be a flexible, scalable, and reliable tool for ensuring AI systems are not only accurate but also transparent, and trustworthy. This framework lays the foundation for more informed, ethical, and responsible deployment of AI systems in critical decision-making processes across various industries. UES thus represents a step forward in the standardization of explain ability metrics, helping bridge the gap between AI model performance and the need for trust in AI-driven applications. The development of the Unified Explain Ability Score (UES) opens several avenues for future research and practical advancements: While our work demonstrated UES in healthcare and finance, there is potential to adapt the metric for other critical domains like autonomous driving, legal decision-making, and defense. Exploring how UES can be tailored to these sectors will strengthen its versatility and reliability across diverse AI applications. Future work could investigate how UES aligns with human interpretability, focusing on how end users (clinicians, financial analysts, etc.) interact with and perceive explanations. This could lead to a user-centric version of UES that incorporates subjective feedback into the metric's calculation. With increasing regulation on AI

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0032
e ISSN: 2584-2854
Volume: 03
Issue:02 February 2025
Page No: 185-193

systems, there is a need to align UES with emerging legal and ethical frameworks. Incorporating UES as part of compliance checks for explain ability in AI systems could support adherence to guidelines from bodies like the EU's AI Act or the FDA's AI guidelines in healthcare.

References:

[1]. Gunning, D., Aha, D. W. "DARPA's Explainable Artificial Intelligence (XAI) Program." AI Magazine, 2019.
[2]. Doshi-Velez, F., Kim, B. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv preprint arXiv:1702.08608, 2017.
[3]. Ribeiro, M. T., Singh, S., Guestrin, C. "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, 2016.
[4]. Lipton, Z. C. "The Mythos of Model Interpretability." arXiv preprint arXiv:1606.03490, 2016.
[5]. Vilone, G., Longo, L. "Explainable Artificial Intelligence: A Systematic Review." arXiv preprint arXiv:2006.00093, 2020.
[6]. Hoffman, R. R., et al. "Metrics for Explainable AI: Challenges and Prospects." IEEE Intelligent Systems, 2018.
[7]. Arya, V., et al. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques." arXiv preprint arXiv:1909.03012, 2019.
[8]. Bhatt, U., et al. "Evaluating and Aggregating Feature-based Model Explanations." Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS), 2020.
[9]. Mittelstadt, B., et al. "Explaining Explanations in AI." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.
[10]. Tomsett, R., et al. "Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems." arXiv preprint arXiv:1806.07552, 2018.
[11]. Lakkaraju, H., et al. "How Do I Fool You? Manipulating User Trust via Misleading Black Box Explanations." Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019.
[12]. Ribeiro, M. T., Singh, S., Guestrin, C. "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, 2016.
[13]. Lundberg, S. M., Lee, S.-I. "A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems, 2017.
[14]. Simonyan, K., Vedaldi, A., Zisserman, A. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." 2013.
[15]. Hoffman, R. R., et al. "Metrics for Explainable AI: Challenges and Prospects." IEEE Intelligent Systems, 2018.
[16]. Vilone, G., Longo, L. "Explainable Artificial Intelligence: A Systematic Review." 2020.
[17]. Van der Lee, C., et al. "Best Practices for the Human Evaluation of Automatically Generated Text." Proceedings of the 12th International Conference on Natural Language Generation, 2019.
[18]. Doshi-Velez, F., Kim, B. "Towards a Rigorous Science of Interpretable Machine Learning." 2017.
[19]. Kindermans, P.-J., et al. "The (Un)reliability of Saliency Methods." 2017.
[20]. Miller, T. "Explanation in Artificial Intelligence: Insights from the Social Sciences." Artificial Intelligence, 2019.
[21]. Arya, V., et al. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explain Ability Techniques".
[22]. Ribeiro, M. T., et al. "Why Should I Trust You? Explaining the Predictions of Any Classifier."

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0032
e ISSN: 2584-2854
Volume: 03
Issue:02 February 2025
Page No: 185-193

[23]. Hoffman, R. R., et al. "Metrics for Explainable AI: Challenges and Prospects."
[24]. Alvarez-Melis, D., Jaakkola, T. "On the Robustness of Interpretability Methods."
[25]. Miller, T. "Explanation in Artificial Intelligence: Insights from the Social Sciences."
[26]. Arya, V., et al. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explain Ability Techniques."