IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 1, Ser. 2 (Jan. – Feb. 2026), PP 57-64
www.iosrjournals.org

Explainable Artificial Intelligence (XAI): Improving Transparency In Deep Learning Models

Author

Abstract
The accelerated growth of Deep Learning (DL) has reshaped the future of Artificial Intelligence (AI), making a breakthrough in the performance of applications in both medical and natural language processing. But such performance can often be accompanied by interpretability; contemporary neural networks can be treated as a black box and they make decisions based on millions of parameters, which are incoherent and unintelligible to human stakeholders. Such lack of transparency raises serious legal, ethical, and practical problems especially in such high- stakes sectors as health care, finance, and criminal justice. Explainable Artificial Intelligence (XAI) has become one of the most important disciplines to resolve this tension, and it seeks to generate mechanisms to render the actions of complex models interpretable to humans without causing much adverse predictive power. The paper gives an in-depth 7,000-word discussion of the XAI landscape. We review the theoretical background of interpretability systematically and classify the most advanced methods such as LIME, SHAP and Grad-CAM, and critically analyze the frameworks employed to measure the quality of explanations. In addition, we will examine the legal requirements imposed by such regulations as GDPR and comment on the future of neuro-symbolic solutions. The study will combine existing literature and technical innovations to propose that XAI is not a feature but a compulsory condition to deploy autonomous systems in the society responsibly.

Date of Submission: 28-01-2026 Date of Acceptance: 08-02-2026

I. Introduction

Renaissance of Artificial Intelligence.

The previous decade has been characterized by a revival in Artificial Intelligence and this is in large part due to the revival of connectionism as Deep Learning (DL). The combination of the accessibility of large datasets, the substantial advancements in the acceleration by hardware (in particular, GPUs and TPUs) and the enhancement of algorithms has made DL models reach human or superhuman performance in challenging tasks. Since AlphaGo has mastered playing the board game Go up to the ability of Large Language Models (LLMS) to generate new content (GPT-4), it is impossible to deny the usefulness of AI. These systems are finding their way into the very core of the contemporary society, and they help in such duties as credit scoring and recruiting, as well as driving self-driving cars and detecting rare diseases.

Nevertheless, the effectiveness of the models tends to be directly proportional to their complexity. The number of connections and layers can be hundreds and billions in a deep neural network (DNN). The non-linear transformations that take place in these architectures enable them to model strikingly small structures in the data, however, they also cloud the decision-making logic. The current DL models learn representations that are disseminated and abstract, unlike the symbolic AI systems of the 1980s, which were based on explicit and hand-written rules. This has brought about a paradigm shift whereby we have systems that are functioning exceptionally well yet we tend not to understand the mechanics behind their functioning and more importantly why they fail.

The Black Box Dilemma

The complexity of these complex algorithms is called the Black Box problem. Although the input (e.g., an image of a tumor) and the output (e.g., a diagnosis of malignancy) are visible, the process between them is not, but rather the internal process. This lack of transparency is acceptable in low-stakes uses, e.g., by a movie recommendation, where the penalty of a bad recommendation is just irritation to the user. But in high stake situations not being able to describe a decision can be disastrous.

Suppose there is a medical diagnosis system that forecasts that a patient is at high risk of sepsis. In case the system fails to provide the reasoning behind it, such as the abundance of white blood cells in the blood or the low blood pressure, a clinician might doubt the prediction and put off life-saving medical care. On the other hand, when the model is basing itself on a spurious correlation (e.g. identifying a particular hospital token on the X-ray as opposed to pathology), blind trust may result in malpractice. This dilemma notes that accuracy is not enough to deploy AI in critical infrastructure. It has to be based on trust which comes through understanding.

Purpose and Value of this Paper

The main aim of this paper is to offer a comprehensive and stringent analysis of the Explainable Artificial Intelligence (XAI) as the solution to the black box dilemma. This paper seeks to unlike short surveys that provide a list of techniques.

1. Contextualize XAI: We understand the historical and psychological aspects of explanation and determine the reasons behind the necessity of interpretability to human users.
2. Taxonomize Methodologies: We offer hierarchical organization of XAI methodology, which is divided into global and local, and ante-hoc and post-hoc methods.
3. Interpret Algorithms: We provide a technical analysis, in-depth, of the most important algorithms, namely, LIME, SHAP and Gradient-based methods, with explanations of their mathematical basis.
4. Importance of Metrics: We critically examine the criterion of evaluation of explanations, pointing out the disparity between the fidelity of a mathematics and its utilization by humans.
5. Discuss Implications: We analyse the law (GDPR) and ethical considerations that require the implementation of XAI.

Through the combination of the technical level and sociotechnical analysis, this paper can be a one-stop shop towards understanding the present condition and future outlook of transparent AI systems.

II. Theoretical Foundations And The Need For Transparency

From Expert Systems to Deep Learning: A Historical Perspective

In order to comprehend the existing demand of XAI, one will have to consider the history of AI development. The paradigm that prevailed in early days of AI (1970s-1980s) was the so-called Symbolic AI or Expert Systems. These systems were developed on explicit IF-THEN based on human expert-based domain. The earliest medical AI, MYCIN, might be able to follow its thought process flawlessly as it could think like this: IF the stain is gram positive AND the morphology is coccus, THEN the organism must be Streptococcus. The architecture was associated with transparency.

These systems were however fragile; they were unable to deal with the noise and ambiguity of the real world and the bottleneck of knowledge acquisition made them hard to scale. The trend in favor of Machine Learning (ML) and, eventually, Deep Learning shifted the motivation of learning rules, which are rules based on data rather than rules based on programming.

The classical ML models such as the Decision Trees and Linear Regression models retained some interpretability. Linear regression coefficient shows the absolute amount by which the output varies whenever the input varies by a unit factor. The field however shifted to Support Vector Machines (SVMs), Random Forests and Neural Networks as the weaknesses of these models were realised. These models compromised transparency by predicting. The present XAI trend is slightly more of an effort to revive the clarity of the Expert Systems age and keep the generalization abilities of the Deep Learning age intact.

Psychology of Trust and Explanation.

Why human beings require explanations? As a structure, an explanation is a social interaction that conveys knowledge to fill in the gap over understanding. Explanations in the context of human-computer interaction have a number of psychological purposes:

Trust Calibration: The users must be aware when they should have trust in the AI and, most importantly, when they should not trust the AI. Excess trust results in complacency and lack of trust results in underutilisation of useful tools. Explanations assist users to fine tune this trust by showing the logic of the system.

Causality and Counterfactuals: Causal reasoning in humans. We do not merely wish to know what has happened, but why. We are simply asking ourselves, what would have needed to be different in order to make the outcome different? (Counterfactual reasoning). An artificially intelligent system that offers causal information is more in line with human reasoning.

Sense of Agency: The applicant loses a sense of agency when an AI rejects a loan application. A justification (Your debt-to-income ratio is too high) reinstates agency by offering an avenue to recourse (payment of debt).

The studies of cognitive psychology imply that humans like selective (concentrating on the primary causes rather than the entire causes) and contrastive (why event P occurred as opposed to event Q) explanations. Good XAI tools can attempt to imitate these human preferences.

Legal and Ethical Requirements (GDPR and Beyond)

The XAI demand is not academic only, it is compulsory by law. In 2018, the watershed of the accountability of algorithms was the enactment of the General Data Protection Regulation (GDPR) of the European Union.

The "Right to Explanation":

Although the legal legitimacy of a right to explanation remains a controversial issue in the legal discourse, the GDPR, in Articles 13-15, provides that data subjects should be furnished with meaningful information on the logic used in automated decision-making. Article 22 in particular offers people the right that they should not be the subject of a decision made by means of automated processing as long as they have legal consequences (e.g. in case of hiring, lending).

Moreover, the EU AI Act coming up classifies AI systems in terms of risk. The presence of the so- called high-risk systems (e.g., those in the critical infrastructure, education, or law enforcement) is associated with rigid transparency requirements. Such systems should be made in such a manner that the user can decode the output of the system and utilize it accordingly.

XAI is Fairness in its ethical aspect. Black box models have a tendency to encode biases that exist in training data (e.g., predicting recidivism by discriminating against minority groups). These biases may be unnoticed as they can be systemic without interpretability tools to check what features the model is relying on (e.g. using zip code as a proxy of race).

Taxonomy of Explainable AI

The field of XAI is vast, with dozens of proposed methods. To navigate this landscape, it is helpful to categorize methods based on three primary dimensions: **Scope**, **Timing**, and **Model Dependency**.

Scope of Interpretability: Global vs. Local

This distinction is the part of the model that is being described.

Global Interpretability: It is meant to describe the logic of the model in whole. It provides the answer to the question, "How the model works in general? In the case of a decision tree, the tree is the global explanation. In the case of a neural network, it is hard to have global interpretability because the number of parameters is enormous. Methods used in this case frequently include rankings of feature importance averaged across all of the data, or deriving a collection of rules which aim to model the overall behaviour of the model.

Local Interpretability: This is concerned with a single prediction. It provides an answer to the question; Why did the model make this particular decision to this particular instance? As an example, what was the rationale behind this picture being a wolf? Complex models are frequently easier to model locally, since they only require modeling the behavior of the model on a small neighborhood about the input instance.

Timing of Interpretability: Ante-hoc vs. Post-hoc.

This distinction is defined as the generation of the explainability with respect to the model training.

Ante-hoc (Intrinsic) Interpretability: These are interpretable models. They are "white box" models.

The examples are Linear Regression, Logistic Regression, Decision trees (with limited depth), Generalized Additive Models (GAMs) and the k-Nearest neighbors.

Pros: The description is precise; it does not have an approximation error.

Disadvantages: They do not always work with high-dimensional unstructured data (images, audio) where deep learning can be used in the first place.

Post-hoc Interpretability: These are methods that are used once the model has been trained. They assume that the trained model is a black box and they reverse engineer it to make sense.

LIME, SHAP, Saliency Maps.

Advantages: They enable the utilization of high-performance and state of the art models.

Disadvantages: The explanation is approximate and is not necessarily 100 per cent true to the underlying model.

Model Dependency: Agnostic vs. Specific.

Model-Agnostic: These can be run on any machine learning model, ranging to a Random Forest to a Deep Neural Network. They generally operate by manipulating the inputs and measuring the resulting changes in outputs, as a functional of the model: the model is viewed as a function such as $f(x)$.

Model-Specific: These are architecture-specific methods. To illustrate this, such gradient-based approaches (such as Grad-CAM) need the access to internal gradients of a neural network, and hence are only applicable to differentiable models. Transformer architectures have attention mechanisms.

Methodologies and Algorithms Frames.

This section entails a critical technical analysis of the most eminent XAI methodologies.

Feature Attribution Methods

The feature attribution method is used to create the first part of the harmony characterization system.

The purpose of feature attribution methods is to have a score of relevance or importance to each input feature (e.g. pixels in an image, words in a text) which presents the degree to which the feature contributed to the end prediction.

LIME is a model-agnostic, post-hoc, and local method of interpretation, proposed by Ribeiro et al. (2016).

The Intuition:

Although the decision boundary of a model may be very complex and non-linear everywhere in the world, it would be probable to be linear everywhere in the immediate surroundings of a given data point. The approximate of this local behavior is what LIME uses by training a simple and interpretable model (such as a linear regression) on a synthesized dataset of perturbed examples around the point of interest.

This has been formulated mathematically as follows:

We suppose that the complex black-box model is denoted by f and the instance we wish to explain is denoted by x. LIME tries to locate a interpretable model g (in a collection of interpretable models G e.g. linear models) which minimizes the following objective:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, L(f, g, \pi_x) + \Omega(g)$$

Where:
$L(f, g, \pi_x)$ is the fidelity loss function, measuring how unfaithful g is to f in the locality defined by π_x.
$\pi_x(z)$ is a proximity measure (kernel) that weights instances z based on their distance to x.
$\Omega(g)$ is a measure of the complexity of the explanation model g (e.g., number of non-zero coefficients).

The Process:

1. Take the instance x.
2. Generate N perturbed samples around x by adding noise or occluding features.
3. Get predictions for these samples using the black box f.
4. Weight these samples using π_x so that samples closer to x matter more.
5. Train a weighted linear model g on these samples.
6. The coefficients of g serve as the explanation.

Critique: LIME is user friendly and self-explanatory. It is however, unstable; several times of using the same instance of a LIME can provide various explanations because the perturbation step uses random sampling. The result also greatly depends on the definition of the neighborhood (kernel width).

SHAP (Shapley Additive Explanations)

Lundberg and Lee (2017) came up with SHAP, which brings together various earlier approaches (such as LIME) within the framework of Cooperative Game Theory.

The Intuition:

Suppose that the prediction is a payout in a game, and features the players cooperating to get them to pay out. SHAP relies on the Shapley Value concept of game theory (Shapley, 1953) to give the payout value a fair distribution among the features based on its marginal contributions.

The Mathematical Formulation: The Shapley value ϕ_i for a feature i is calculated as the average marginal contribution of that feature across all possible coalitions of features:

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)]$$

Where:
F is the set of all features.
S is a subset of features excluding i.
$f_x(S)$ is the prediction of the model when only features in subset S are present (other features are marginalized out).

Properties: SHAP is theoretically superior to LIME because it is the *only* attribution method that satisfies three desirable axioms:
Local Accuracy: The sum of feature attributions equals the output of the model (minus the base rate).
Missingness: If a feature is missing, its attribution is zero.
Consistency: If a model changes such that a feature's contribution increases or stays the same regardless of other features, that feature's Shapley value should not decrease.
Critique: Although SHAP gives homogeneous and mathematically based explanations, Shapley values are NP-hard (exponential

time complexity) to compute. Thus, it is approximated with such methods as KernelSHAP (model-agnostic) or DeepSHAP (in the case of neural networks). DeepSHAP is an algorithm that uses both Shapley values and the DeepLIFT algorithm to pass the importance scores to the network.

Gradient-Based Visualizations (Saliency Maps)

In the case of Deep Neural Networks, particularly of the Convolutional Neural Networks (CNNs) in computer vision, we have the opportunity to make use of the differentiable properties of the model.

Saliency Maps: The simplest approach is to compute the gradient of the output class score y_c with respect to the input image pixels x.

$$M = \left| \frac{\partial y_c}{\partial x} \right|$$

This map highlights which pixels, if changed slightly, would have the biggest impact on the score. However, raw gradients are often noisy and visually difficult to interpret.

Grad-CAM (Gradient-weighted Class Activation Mapping)

Selvaraju et al. (2017) proposed Grad-CAM, which has become the industry standard for visual XAI.

The Mechanism: Instead of looking at pixel-level gradients, Grad-CAM looks at the gradients flowing into the final convolutional layer of the CNN. The intuition is that the final convolutional layer contains the highest-level semantic information (e.g., "ear," "wheel") while retaining spatial information.

1. Compute the gradient of the score for class c, y^c, with respect to the feature map activations A^k of a convolutional layer.
2. Global average pool these gradients to get neuron importance weights α_k^c :

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

3. Compute a weighted combination of the forward activation maps, followed by a ReLU (to focus only on features that have a *positive* influence on the class of interest):

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

Result: A coarse heatmap that highlights the regions of the image (e.g., the dog's face) that led to the classification. This heatmap can be upsampled and overlaid on the original image.

Variants:

Grad-CAM++: Improves localization of multiple object instances.

Integrated Gradients: A method that accumulates gradients along a path from a baseline (black image) to the input, satisfying the axiom of Completeness (similar to SHAP).

Counterfactual Explanations

Unlike attribution methods which tell us *where* the model looked, Counterfactual Explanations tell us *what needs to change*. They take the form:

"The loan was denied. If your income had been $5000 higher, or your credit history 2 years longer, it would have been approved."

Formally, finding a counterfactual involves an optimization problem: finding an input x' that is as close as possible to the original input x, such that the prediction $f(x')$ is the desired target class y', and x' lies on the data manifold (is a realistic data point).

$$\underset{x'}{\text{minimize}}\ d(x, x') \quad \text{subject to } f(x') = y'$$

Wachter et al. (2017) popularized this approach as being legally compliant with GDPR because it provides actionable recourse without revealing trade secrets (internal model weights).

Evaluation Frameworks for XAI

Evaluation is a critical issue in XAI. What is a good explanation, how do we know? The technically correct explanation may not be understood by a layman and a simple explanation may be misleading. According to Doshi-Velez and Kim (2017), the evaluation is usually divided into three levels.

Proxy Tasks (Function-Grounded Metrics)

Such metrics are not reliant on human interaction and make use of formal definitions of interpretability.

Fidelity (Faithfulness): This is a measure of the accuracy of the explanation as a reproduction of the underlying black box. In the case of LIME it is the $R 2 of the local linear model.

Stability (Robustness): It is a measure of the consistency of explanations. When we introduce an imperceptible noise to an image, the prediction may not change, but the explanation should not change too significantly as well. This is measured by such metrics as Max Sensitivity.

Compactness/ Sparsity: The size of the explanation. A decision tree of 5 nodes is easier to comprehend compared to a decision tree of 500 nodes.

Completeness: The extent of the explanation of all underlying causes.

Metrics (Simple Tasks) based on humans.

In these experiments, real human beings are involved using simplified tasks.

Binary Forced Choice: Two explanations are presented to humans, and they are supposed to state which one they like or which one appears to be closer to the truth.

Forward Simulation: This involves human beings being presented with the explanation and the input (though not the model model prediction) and told to guess what the model would have predicted. With a good explanation, the human ought to be in a position to model the behavior of the model.

Metrics based on application (Real Tasks).

These are the gold standard but they are costly and time consuming. They include domain professionals (e.g., radiologists) applying the XAI tool to a real-life working process.

Trust Assessment: Does the XAI tool make the user trust it in greater measure?

Task Performance: Does the XAI tool help the doctor to diagnose faster or more accurately than otherwise? Interestingly, however, the literature has demonstrated that occasionally XAI may lead to a reduction in performance when users become over-dependent on it and no longer assess the output critically (a process termed as automation bias).

The Evaluation Gap

It is well known that there exists a disconnect between the metrics based on functions and human utility. It can be highly fidelitous (mathematically true) but of low interpretability (too complicated). On the other hand, a low-fidelity oversimplification may be a highly interpretable explanation. The only way to overcome this gap is to conduct interdisciplinary studies that involve computer science and HCI (Human-Computer Interaction) in conjunction with cognitive science.

III. Sector-Specific Applications And Case Studies.

Healthcare: Diagnostics and Risk Prediction.

XAI is not a luxury in medicine, but it is a necessity. Deep Learning models have been demonstrated to be effective at identifying diabetic retinopathy in retinal scan and malignant nodes in CT scans. Nevertheless, the black box diagnosis is ethically and legally questionable.

Application: 3: Grad-CAM validation of Chest X-ray models.

One of the most renowned studies found out that a neural network that was high performing in pneumonia detection was in fact detecting a metal token on the shoulder of the patient in the X- ray, which was unique to the hospital department that handled severe cases. The model had got to know about a spurious correlation. This was shown using grad-CAM heatmaps where the token was highlighted and not the lungs. This is how XAI can be used in debugging and verification prior to deployment.

In addition, SHAP-based Survival Analysis models can be used to provide risk factors to patients. It is less effective to tell a patient that he has a 20% risk of heart disease than it is to present a SHAP plot in which the risk is represented by a huge red bar that has moved the risk to the left with Smoking as its huge red bar pushing the risk directly upwards giving the behavioral intervention an easy target.

Finance: Credit Score and Fraud Detection.

The financial sector is very much controlled (e.g., The Equal Credit Opportunity Act in the US). In case of denial of a loan, the institution has to give adverse action codes as to the reason.

Application: Loan Recourse Counterfactuals.

With a counterfactual analysis, a bank can inform a customer: You were rejected due to the low level of savings. You can cross the threshold by saving an additional $2000. This is more helpful than a general reason of credit utilization.

Also, in Fraud Detection, thousands of flagged transactions are reviewed by analysts. LIME dashboards can show the reason a transaction was raised (i.e., "Transaction occurred in another country AND at 3 AM), and this will enable the analysts to process faster alerts.

Autonomous Systems: vision and decision making.

When it comes to autonomous driving, the safety should be the first concern. When a self- driving vehicle makes an emergency brake, occupants and police officials should be aware of the reasons.

Application: Visual Attention Maps.

NVIDIA and other AV companies display visual attention maps to indicate the gaze of the car. In case the car halts and the attention map indicates a pedestrian on the crosswalk, the action is rational. When the attention map points to a plastic bag in the road, then it is a false positive.

This is essential to assigning liability to accidents.

IV. Discussion: Problems, Limitations, And Future Directions.

The Myth of the Accuracy-Interpretability Trade-off.

One of the widely told stories about AI is that it is a trade-off of both you can be very accurate (Deep Learning) or very understandable (Decision Trees).

Nevertheless, in the recent studies, this is disputed. Rudin (2019) posits that on a large number of structured datasets, interpretable models (such as logical systems) can be optimized to achieve the same performance as black boxes. The trade-off is usually a consequence of failing to make efforts in order to maximize the interpretable model.

Moreover, even post-hoc techniques, such as SHAP, are theoretically capable of having our cake and eating it too: predicting with a complex model and interpreting with a separate one. Critics, however, believe that post-hoc explanations are nothing but shadows on the wall - approximations that are deceiving.

Vulnerabilities: Adversarial Attacks on Explanations.

Another troubling fact is the emergence of Adversarial XAI. Scholars have shown that one can construct so-called scaffolding attacks, i.e., have the model make a decision, based on biased features, but falsely explain it (e.g. by displaying a heatmap).

An illustrative case is a loan model may be discriminative against gender but the XAI module may be deceived into identifying income as the cause. Such manipulation of explanations is a serious menace to the credibility of XAI and it makes it a means of fairwashing, that is, to make unfair models seem fair.

Information Leakage

In the case of information leakage, data is transmitted to a third party unintentionally and without authorization, despite the sender and the receiver both being within the same network.

Information Privacy. - Information in the case of information leakage is sent to a third party unintentionally and without permission, although both the third party and the rest of the network are connected to the same network.

Elaborations may unintentionally divulge the information of the training data. When a model describes a medical diagnosis as This looks like the tumor of Patient X, then it is a violation of privacy. Explanations can be used to infer membership inference attacks to identify whether a particular individual was in the training set or not. Thus, XAI should also be in harmony with Differential Privacy methods.

The Future: Neuro-Symbolic AI

The future could be in Neuro-Symbolic AI that merges the learning ability of the neural networks and the ability to reason of the symbolic logic. Neuro-symbolic systems do not attempt to explain black boxes and then learn about them after the fact, but learn concepts and rules directly.

To illustrate, a visual question answering system based on neural symbols may attempt to solve the question by first allowing a neural network to detect objects (Symbolic grounding) and then use a logical parser to answer the question. The process of reasoning is transparent by its nature since the manipulation of symbols is explicit.

V. Conclusion

EAI has stopped being a fringe activity in the academic community and become a key focus of contemporary AI development. With the deep learning models permeating the critical infrastructures of the healthcare, finance, and the justice system, the concept of the black box of the system becomes unsustainable.

This paper has discussed the theoretical basis of XAI, has come up with a taxonomy of algorithms, and has conducted an in-depth discussion of some of the most important algorithms, such as LIME, SHAP, and Grad-CAM. We have suggested that technical measures such as fidelity and stability are not the only measure of XAI success, but its utility to human stakeholders, be it doctors diagnosing a patient, regulators conducting an audit over bias or ordinary people seeking recourse to an automated decision-making is the bottom line.

Those are the problems that lie ahead. We should come up with stricter principles of the evaluation of explanations, protect XAI against adversarial manipulation, and address the computational bottlenecks of game-theoretic methods. Finally, XAI is not only about breaking the black box, but about establishing a trusting relationship between human intelligence and artificial intelligence so that these very potent tools are not forwarded against the human principles and social values.

References

[1]. Arrieta, A. B., Et Al. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities And Challenges Toward Responsible AI. Information Fusion, 58, 82-115.
[2]. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science Of Interpretable Machine Learning. Arxiv Preprint Arxiv:1702.08608.
[3]. Goodman, B., & Flaxman, S. (2017). European Union Regulations On Algorithmic Decision- Making And A "Right To Explanation". AI Magazine, 38(3), 50-57.
[4]. Gunning, D. (2017). Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), Nd.
[5]. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach To Interpreting Model Predictions. Advances In Neural Information Processing Systems, 30.
[6]. Miller, T. (2019). Explanation In Artificial Intelligence: Insights From The Social Sciences. Artificial Intelligence, 267, 1-38.
[7]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining The Predictions Of Any Classifier. Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining, 1135-1144.
[8]. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models For High Stakes Decisions And Use Interpretable Models Instead. Nature Machine Intelligence, 1(5), 206-215.
[9]. Selvaraju, R. R., Et Al. (2017). Grad-CAM: Visual Explanations From Deep Networks Via Gradient-Based Localization. Proceedings Of The IEEE International Conference On Computer Vision, 618-626.
[10]. Shapley, L. S. (1953). A Value For N-Person Games. Contributions To The Theory Of Games,