Formulating a Theoretical Framework for Deep Learning Model Comprehensibility

Jialong Jiang
School of Software, Jiangxi Normal University, Nanchang,Jiangxi 330000, China

Abstract—This Deep learning models have achieved remarkable success across various fields. However, their 'black box' nature has become a fundamental issue. Explainability aims to bridge this gap by providing insights into the decision - making processes of models. This paper delves into the theoretical foundations of explainability in deep learning, focusing on mathematical and conceptual aspects.We examine the limitations of current explainability approaches and discuss how interdisciplinary methodologies can enhance our understanding of deep learning systems. Moreover, we explore the potential of integrating explainability with robustness, fairness, and generalization to create more reliable AI systems.The paper also highlights several challenges, such as the trade - off between interpretability and predictive power, the scalability of explainability methods, and the lack of standard evaluation metrics. Furthermore, we propose novel research directions, including topological analysis, causal reasoning, and probabilistic explainability models.Particular attention is given to the role of human cognition, decision - theoretic frameworks, and the use of explainability as a tool to improve the reliability of deep learning models in high - stakes scenarios. We also investigate how explainability techniques can enhance the deployment and optimization of deep learning models in real - world environments, ensuring their ethical and practical applications.This work aims to provide a comprehensive framework for improving the transparency, interpretability, and accountability of AI - driven decision - making systems.

Index Terms—Explainability in Deep Learning, Interpretability-Performance Trade-off, AI Robustness and Fairness, Causal Reasoning, Human-Centered Explainability

I. INTRODUCTION

The rapid advancement of deep learning has driven its application across various fields, including image recognition, natural language processing, medical diagnostics, and financial forecasting. These models have shown exceptional predictive power. However, their growing complexity and dependence on large - scale datasets have also brought new transparency and trust challenges. Users, stakeholders, and regulatory bodies require clear explanations of how AI systems make decisions, particularly when these decisions affect people's lives. As deep learning continues to shape the technological landscape, explainability has become a key focus for researchers and practitioners.

A primary reason for enhancing explainability is to ensure accountability in decision - making. In high - stakes areas like criminal justice and autonomous driving, AI models must provide justifiable and interpretable decisions. Without proper transparency, deep learning systems may spread biases, strengthen discriminatory patterns, or make untraceable errors. Explainability is vital for reducing these risks by offering insights into model behavior, identifying biases, and ensuring ethical and responsible AI - driven decisions.

Moreover, explainability is crucial in AI development. Engineers and data scientists need clear explanations to diagnose errors, optimize model architectures, and boost generalization. Debugging complex deep learning systems without interpretability tools is like dealing with a black box, where even small changes in training data or hyperparameters can lead to unpredictable model behavior. By using explainability techniques, researchers can better understand neural network representations, track information flow between layers, and design more robust architectures to resist adversarial attacks[1].

The explainability debate is also complicated by the varying interpretability needs of different stakeholders. For example, medical experts using AI diagnostic tools may need different explanations than laypersons receiving loan approval decisions from financial AI systems. Thus, explainability isn't a one - size - fits - all solution. It requires interdisciplinary collaboration among AI researchers, legal experts, ethicists, and cognitive scientists to develop user - centered interpretability frameworks that consider different complexity levels, granularity, and audience requirements[2,3].

Another important factor is the trade - off between explainability and model performance. Some highly accurate deep learning models, like large - scale transformer architectures, are among the least interpretable due to their complex attention mechanisms and billions of parameters. Researchers must balance these competing objectives, aiming to create models that are both highly accurate and capable of providing meaningful explanations. Recent progress in self - explainable AI models, hybrid neuro - symbolic approaches, and modular architectures shows promise in addressing this challenge.

The rest of this paper is organized as follows. Section 2

This work was not supported by any specific funding agency.

Jiang Jialong is with the School of Software, Jiangxi Normal University, Nanchang,Jiangxi,China 330000 (e-mail: jialong@jxnu.edu.cn).

looks into the theoretical foundations of explainability, exploring the key mathematical and conceptual frameworks that support interpretability in deep learning. Section 3 examines major challenges and open research questions, such as the scalability of explainability methods and the transparency - performance trade - off. Section 4 outlines future research directions, highlighting emerging trends like causal explainability, real - time interpretability techniques, and fairness - aware AI models. Finally, Section 5 concludes with a discussion on the broader implications of explainability for the future of artificial intelligence[4,5].

The widespread adoption of deep learning models has transformed numerous fields, from healthcare and finance to autonomous systems and natural language processing. However, these models remain largely opaque, making it difficult for practitioners, regulators, and end - users to understand how decisions are made. This lack of transparency poses significant challenges regarding accountability, fairness, and trustworthiness, especially in high - risk applications where model decisions can have serious consequences[6].

The demand for explainability in deep learning comes from several factors. First, regulatory frameworks like the General Data Protection Regulation (GDPR) emphasize transparency in automated decision - making. Second, biases in AI models have raised concerns about fairness and ethics, making interpretable approaches more necessary. Third, the vulnerability of deep learning models to adversarial attacks highlights the need to better understand decision boundaries and robustness properties. Lastly, as AI systems become more integrated into human - centric applications, their behavior must align with human reasoning and domain knowledge. Addressing these issues requires a multidisciplinary approach that combines insights from computer science, cognitive psychology, philosophy, and ethics[7].

Explainability is often mentioned alongside interpretability, but they differ in scope and method. Interpretability refers to how understandable a model is to humans, while explainability focuses on providing a post - hoc or intrinsic understanding of how and why a model makes a specific decision. Various techniques, such as feature attribution, model distillation, symbolic reasoning, and causal inference, have been developed to enhance explainability. Despite progress, creating universally applicable and reliable explainability methods remains challenging.

This paper offers an in - depth analysis of the theoretical foundations of explainability in deep learning. We explore the mathematical and conceptual basis of existing techniques, discuss their limitations, and investigate future directions for improving model transparency. By incorporating perspectives from information theory, geometry, and causality, we aim to provide a comprehensive framework for understanding deep learning system explainability. We also emphasize the practical importance of explainability in AI - driven solution deployment[8], ethical compliance, and trust - building. Furthermore, we examine real - world XAI applications across different sectors, showing how better interpretability can improve model adoption, debugging, and risk assessment.

II. Theoretical Foundations of Explainability

A. Information Theory

Information theory is pivotal in enhancing the explainability of deep learning models. The information bottleneck principle posits that these models compress input data into essential features for prediction. While this process reduces redundancy, it can also compromise the interpretability of learned representations. By studying how information is preserved or lost across a network, researchers can uncover insights into the model's decision-making mechanisms. Information flow analysis further aids in designing models that strike a balance between compression and interpretability, ensuring critical features are retained during training. A range of methods, including mutual information estimation, entropy analysis, and rate-distortion theory, offers quantitative tools to assess explainability in deep networks. Moreover, information-theoretic approaches are instrumental in understanding generalization bounds, shedding light on how well a model's learned representations apply to unseen data[9,10].

B. Geometric and Topological Analysis

The geometric and topological analysis of neural networks provides another powerful framework for understanding their decision-making processes. Neural networks essentially transform input data via nonlinear operations, embedding them into high-dimensional manifolds. Persistent homology, Riemannian geometry, and algebraic topology are some of the tools proposed to study how these transformations influence decision boundaries and feature separability. By examining the geometric structure of learned representations, researchers can gain valuable insights into the inner workings of deep networks and their generalization properties. Manifold learning and curvature analysis enable a more structured approach to interpreting feature space evolution within neural networks. Understanding geometric disentanglement in latent spaces can reveal the factors contributing to model decisions, thereby enhancing interpretability. Topological data analysis (TDA) has also proven useful in characterizing the robustness of deep networks by analyzing the stability of learned features under perturbations. Furthermore, advances in deep metric learning and contrastive representation learning have led to a more structured understanding of latent space organization in deep networks[11].

C. Causal Inference

Causal inference is crucial for distinguishing correlation from causation in deep learning explainability. Traditional machine learning models often rely on correlational patterns in data, which can lead to misleading explanations. Causal inference techniques like counterfactual reasoning, structural causal models (SCMs), and do-calculus offer a more rigorous framework for understanding why a model makes specific decisions. Integrating causal reasoning into deep learning architectures can result in more reliable and interpretable models that align with human intuition[12]. Causal discovery methods can uncover hidden dependencies in neural networks, improving their robustness and trustworthiness. Recent

developments in causal representation learning enable the incorporation of causal knowledge into deep learning, fostering more transparent and generalizable AI models. Causal disentanglement techniques allow for the isolation of independent generative factors, ensuring that learned representations reflect meaningful real-world relationships and thus enhancing interpretability. Additionally, combining causal modeling with adversarial robustness techniques helps maintain model explainability even under adversarial conditions[13].

D. Symbolic AI and Neuro-Symbolic Integration

Symbolic AI and neuro-symbolic integration add another dimension to explainability. Symbolic reasoning involves explicit rule-based logic and has long been regarded as interpretable. In contrast, neural networks are more data-driven but less transparent. Hybrid models combining symbolic reasoning with deep learning show great potential for creating inherently interpretable AI systems. Neuro-symbolic approaches merge the expressiveness of neural networks with the explicit reasoning capabilities of symbolic systems, making AI decisions more comprehensible. These models are especially useful in fields requiring strong reasoning capabilities, such as healthcare, finance, and legal applications. Differentiable programming advancements have enabled seamless integration between symbolic logic and deep networks, resulting in end-to-end trainable neuro-symbolic models that enhance interpretability without sacrificing learning efficiency. The emergence of large-scale neuro-symbolic architectures trained on extensive knowledge bases further strengthens AI models' ability to provide structured and interpretable decision-making processes[14].

Probabilistic modeling contributes to explainability by quantifying uncertainty in predictions. Bayesian deep learning methods provide principled ways to capture model confidence and epistemic uncertainty. Knowing when a model is uncertain about its predictions can significantly improve transparency and trust in AI systems. Probabilistic graphical models, such as Bayesian networks and Markov random fields, further clarify the dependencies among features and model outputs. Combining probabilistic reasoning with deep learning also enhances robustness in real-world deployment, particularly in safety-critical applications where uncertainty must be accounted for. The integration of approximate inference techniques like variational inference and Markov Chain Monte Carlo (MCMC) enables deep models to explicitly represent uncertainty while keeping computational efficiency[15,16].

These theoretical foundations collectively underpin explainability in deep learning. By integrating these concepts, researchers can develop AI models that are both powerful and transparent, ensuring ethical and accountable deployment.

III. Challenges and Open Questions

A. Trade-off Between Accuracy and Transparency

The quest for the optimal balance between model performance and interpretability remains a significant challenge. Complex models like deep neural networks often excel in performance but operate as "black boxes." This trade-off is particularly pronounced in high - stakes fields such as healthcare, finance, and autonomous driving[17,18].

Simplification vs. Fidelity: Techniques like model distillation and attention mechanisms aim to simplify decision - making but may lose critical nuances. Researchers seek ways to design surrogates that accurately represent the decision process without oversimplification.

Algorithmic Trade - offs: Transparent models like decision trees may lack the power of deep learning models. Developing hybrid approaches that combine high accuracy with intrinsic interpretability is an open question. Architectural designs with interpretable modules show promise, but their generalizability across tasks remains unexplored.

Domain - Specific Requirements: Different fields have unique needs. For example, medicine requires clear reasoning for each prediction. The challenge lies in aligning technical interpretability with regulatory and ethical standards while maintaining performance.

B. Scalability of Explainability Techniques

As deep learning models grow, scalability becomes a major concern for explainability methods. Many current techniques are computationally intensive, limiting their practical use in large - scale models or real - time systems.

Computational Complexity: Techniques like feature attribution and saliency maps often require multiple backward passes. For large models, this overhead can be prohibitive in production environments. Optimizing these methods for efficiency without sacrificing explanation quality is crucial.

Modular and Adaptive Architectures: Developing modular frameworks that adapt to different computational budgets and model complexities is promising. Techniques that allocate resources dynamically may offer efficiency, but questions about stability and integration with existing architectures remain[19].

Real - Time Constraints: In dynamic environments like autonomous systems, instantaneous explanations are needed. Generating accurate explanations on - the - fly, especially with evolving models, requires novel methods that meet latency requirements without reducing interpretability[20].

C. Human-Centered Evaluation and Usability

The ultimate goal of explainability is to enhance human understanding and trust. However, many approaches focus on mathematical or computational measures, neglecting the human factor.

User Studies and Psychometric Assessments: Developing evaluation methods that measure explanation effectiveness for the intended audience is critical. Quantitative metrics may not correlate with human comprehension, so user studies assessing interpretability from a cognitive perspective are needed.

Cognitive Load and Information Overload: Balancing detail and clarity is challenging. Overly technical or simplistic explanations can both be problematic. Adaptive systems that personalize content based on user feedback and expertise may provide a solution[21].Context and Relevance: Explanation

effectiveness is context - dependent. Integrating domain - specific constraints into explainability methods requires an interdisciplinary approach combining HCI, cognitive science, and domain expertise.

Transparency vs. Interpretability Trade - offs: Revealing more model details can sometimes cause confusion. Determining the optimal detail level for different contexts is key. Layered explanations offering summaries with optional details may help, but ensuring their coherence across user groups remains an issue.

D.Fairness, Bias, and Ethical Considerations

Explainability intersects with fairness, bias, and ethics in AI systems. Biased explanations can reinforce inequalities and misrepresent decision - making.

Bias in Explanations: Explanation - generation methods can perpetuate biases from training data or models. For example, feature attribution methods might highlight features correlated with sensitive attributes. Ensuring fair and unbiased explanations is a critical research direction.

Ethical Implications: Transparent AI can enhance accountability but raises privacy and misuse concerns. Balancing transparency with privacy protection is essential. Future work must provide meaningful explanations without compromising confidentiality[22].

Regulatory and Legal Challenges: Increasing regulatory scrutiny on AI systems, especially in finance and healthcare, makes ensuring model compliance with legal standards imperative. Integration of explainability into certification frameworks presents challenges and opportunities. Collaboration between researchers and policymakers is needed to develop legally robust standards.

Cross - Cultural and Social Considerations: Interpretability can vary across cultures. Future research should explore how sociocultural factors influence AI explanation perception and develop globally applicable methods.

Adversarial Robustness and Security of Explanations

Ensuring explanation methods are robust against adversarial attacks is an emerging challenge.

Vulnerability to Adversarial Manipulations: Many explanation techniques are sensitive to input perturbations. Adversaries could exploit this to generate misleading interpretations, undermining AI system trust. Developing resilient explanation methods is crucial.

Defense Strategies: Researchers explore combining adversarial training with explainability objectives. However, this interplay may reduce model performance or limit explanation scope[23].

Integration into Verification Processes: Incorporating explainability into model certification and verification is essential for high - stakes applications. Developing standards to evaluate both predictive performance and explanation stability poses technical challenges[24].

E.Interdisciplinary and Theoretical Open Questions

Beyond technical challenges, several theoretical and interdisciplinary questions remain open.

Unified Theoretical Frameworks: Current frameworks often operate in isolation. A unified theory integrating them would facilitate coherent explainability method development.

Metrics and Evaluation Standards: Standardized metrics for explanation quality are needed. Existing metrics may not capture all interpretability aspects. Developing universal evaluation standards is an open question.

Integration with Emerging AI Paradigms: New AI techniques like reinforcement learning pose challenges for generating interpretable explanations. Different methodologies are needed for these paradigms compared to static supervised models.

Scalability of Theoretical Approaches: Scaling theoretical insights to large - scale models is challenging. Bridging the theory - practice gap is essential for translating insights into practical tools.

Interplay Between Explainability and Other AI Properties: The interaction between explainability and other AI properties like fairness and robustness requires understanding. A multidisciplinary approach is needed to build holistic AI systems[25].

F.Future Research Directions and Open Questions

To address these challenges, several promising research directions are emerging:

Hybrid Models: Combining transparent components with high - performing black - box models may offer a middle ground. Research into hybrid models and multi - modal explanations could benefit both performance and understanding.

Adaptive and Personalized Explanations: One - size - fits - all explanations may not work for diverse user groups. Future research could focus on adaptive systems adjusting detail levels based on user expertise, context, and cognitive load[26].

Standardization Efforts: Developing industry - wide benchmarks and standardized evaluation protocols is essential for objective method comparisons and best practice adoption.

Interdisciplinary Collaboration: Solving open questions in explainability requires collaboration among computer scientists, domain experts, ethicists, and policymakers. Initiatives promoting interdisciplinary research are key to developing robust and socially acceptable explanations[27,28].

In summary, explainability challenges are multifaceted, ranging from technical issues like scalability and adversarial robustness to human - centered concerns such as fairness, cognitive usability, and regulatory compliance. These challenges offer many research avenues, driving the field forward. Addressing them can make AI systems more transparent and build the trust needed for their responsible societal deployment.

IV.Future Directions

The future of explainability in deep learning envisions a shift from fragmented, post - hoc methods to integrated, inherently transparent models. This change is driven by the need for models that achieve high predictive performance while offering clear insights into their decision - making processes. Such transparency enhances trust and accountability across

applications.

A key focus is the development of self - explainable architectures. Unlike traditional methods that rely on external techniques to interpret black - box models, self - explainable models have built - in transparency mechanisms. For example, some architectures include interpretable layers that generate explanations alongside predictions. This could involve embedding prototype - based components or specialized attention mechanisms that intuitively highlight critical features. By having models articulate their reasoning during inference, researchers aim to bridge the gap between model performance and human interpretability, ensuring each decision comes with a comprehensible rationale.

Explainability - driven optimization is another promising direction. Traditionally, models have been optimized based on performance metrics. However, interpretability should be a primary training objective. By incorporating explainability into optimization through regularization terms that promote feature sparsity or disentanglement, models can develop effective yet interpretable internal representations. This involves creating new loss functions that balance accuracy and clear explanations. The optimization process thus becomes a dual pursuit of maximizing performance while ensuring transparency.

Integrating adversarial robustness with explainability is also crucial. As models become more prevalent in high - stakes environments, their vulnerability to adversarial attacks poses risks to both prediction accuracy and explanation reliability. Recent research explores methods to ensure explanation stability under adversarial conditions. This means extending adversarial training techniques so models are robust against input perturbations and maintain consistent explanations. Algorithms that jointly optimize for robustness and interpretability are essential for applications where understanding decision - making is as important as the decision itself.

The evolution of interactive and adaptive explanation systems represents another significant frontier. The traditional one - size - fits - all approach is being replaced by systems that tailor outputs to individual users' needs and expertise. For instance, in clinical settings, a diagnostic model might provide a high - level summary for general practitioners and detailed explanations for specialists. These adaptive systems leverage advances in natural language processing and user interface design, enabling real - time interactions between the model and users. By incorporating feedback loops and context - aware algorithms, they continuously refine explanations, enhancing user comprehension and satisfaction. This shift improves AI usability and builds trust by ensuring explanations are relevant and easily understood by diverse audiences.

Standardized evaluation metrics and benchmarks for explainability are also critical. The current variety of evaluation methods, focusing on aspects like fidelity and consistency, lacks universally accepted standards. This makes objective comparison of approaches challenging. Future research must develop comprehensive evaluation frameworks that consider multiple dimensions of explainability. Such standards would facilitate fair comparisons and guide new model design, ensuring they meet transparency and reliability criteria. Collaborative efforts among academia, industry, and regulatory bodies are essential to define these benchmarks and drive the adoption of best practices in AI deployment.

Ethical, legal, and social considerations are increasingly central to the future of explainability. As AI systems are deployed in sensitive domains, ensuring they operate transparently and fairly is paramount. Transparent models can expose biases and prevent discrimination, but they must also be designed with privacy and security in mind. Researchers are now exploring frameworks that embed ethical guidelines into AI systems.

V. Conclusion

In conclusion, the journey toward developing transparent and interpretable deep learning models has revealed both promising avenues and formidable challenges. Our exploration of the theoretical foundations—including information theory, geometric and topological analysis, causal inference, symbolic AI, and probabilistic modeling—has underscored the complexity inherent in balancing model performance with interpretability. These frameworks offer a robust lens through which we can understand the inner workings of neural networks, yet they also highlight the intricate trade-offs that designers face.

The challenges discussed in this paper are multifaceted. On one hand, there is a fundamental trade-off between achieving high accuracy and maintaining transparency. As models become increasingly complex, ensuring that they remain comprehensible to users becomes a daunting task. Current methods like model distillation and attention-based explanations provide valuable insights, but they often fall short of capturing the full complexity of deep learning systems, particularly in high-stakes applications. On the other hand, scalability presents another critical hurdle. Many explainability techniques, especially post-hoc methods, struggle with the computational demands imposed by large-scale models, limiting their practical deployment in dynamic environments.

Moreover, human-centered evaluation of explainability continues to be an essential yet underexplored area. The ultimate goal is to deliver explanations that are not only mathematically robust but also intuitively understandable by diverse user groups. This requires a convergence of research across technical domains, human-computer interaction, and cognitive psychology. Additionally, ensuring fairness, mitigating biases, and enhancing adversarial robustness remain significant challenges. These factors are critical for the deployment of AI systems that are both ethical and reliable.

Looking ahead, the future of explainability lies in the integration of interpretability into every stage of model development—from design and training to evaluation and deployment. Self-explainable architectures and explainability-driven optimization offer promising strategies for creating models that are inherently transparent. At the same time, advances in adversarial robustness and interactive explanation systems are likely to play a key role in enhancing user trust and facilitating real-world adoption.

Ultimately, the pursuit of explainability is not solely a

technical endeavor; it is also a commitment to building AI systems that align with ethical standards and societal values. By continuing to push the boundaries of our understanding and bridging the gap between complex models and human insight, we can pave the way for AI systems that are as accountable as they are innovative. The ongoing research and collaborative efforts in this field hold great promise for a future where AI not only performs exceptionally well but does so in a manner that is transparent, trustworthy, and socially responsible.

REFERENCES

[1] Alvarez-Melis, D., Jaakkola, T. S., 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. Advances in Neural Information Processing Systems 31 (NeurIPS 2018).

[2] Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. _Entropy_ (_Basel_, _Switzerland_), 23(1): 18.

[3] Lipton, Z. C., 2018. In Machine Learning, the Concept of Interpretability is Both Important and Slippery. _Queue_, 16: 28.

[4] Zeng, C. Y., Yan, K., Wang, Z. F., et al., 2021. Survey of Interpretability Research on Deep Learning Models. _Computer Engineering and Applications_, 57(8): 1-9 (in Chinese with English abstract).

[5] Zhang, Y., Tino, P., Leonardis, A., et al., 2021. A Survey on Neural Network Interpretability. _IEEE Transactions on Emerging Topics in Computational Intelligence_, 5: 1-17.

[6] Zhao, H. B., Liu, R., Liu, Y. H., et al., 2022. Research on Classification and Identification of Mine Microseismic Signals Based on Deep Learning Method. _Journal of Mining Science and Technology_, 7(2): 166-174 (in Chinese with English abstract).

[7] Zhou, T., Wu, W., Peng, L., et al., 2022. Evaluation of Urban Bus Service Reliability on Variable Time Horizons Using a Hybrid Deep Learning Method. _Reliability Engineering and System Safety_, 217(3): 108090.

[8] Zhu, L., Huang, L. H., Fan, L. Y., et al., 2020. Landslide Susceptibility Prediction Modeling Based on Remote Sensing and a Novel Deep Learning Algorithm of a Cascade-Parallel Recurrent Neural Network. _Sensors_ (_Basel_, _Switzerland_), 20(6): 1576.

[9] Cheng, K. Y., Wang, N., Shi, W. X., et al., 2020. Research Progress on the Interpretability of Deep Learning. _Journal of Computer Research and Development_, 57(6): 1208-1217.

[10] Fang, R. K., Liu, Y. H., Su, Y. C., et al., 2021. Landslide Susceptibility Early Warning Model Based on Logistic Regression in Qingchuan County, Sichuan. _Hydrogeology and Engineering Geology_, 48(1): 181-187.

[11] Feng, X., Wang, Y., Liu, Y., et al., 2022. Assessment of Landslide Susceptibility Considering the Controlling Mechanism of Weak Interlayers and Their Spatial Uncertainty: A Case Study in Iron Peak, Wanzhou District. _Geological Science and Technology Information_, 41(2): 254-266.

[12] Hu, T., Fan, X., Wang, S., et al., 2020. Landslide Susceptibility Assessment of Sinan County Based on Logistic Regression Model and 3S Technology. _Geological Science and Technology Information_, 39(2): 113-121.

[13] Huang, F. M., Ye, Z., Yao, C., et al., 2020. Uncertainties of Landslide Susceptibility Prediction: Different Attribute Interval Divisions of Environmental Factors and Different Data-Driven Models. _Earth Science_, 45(12): 4535-4549.

[14] Li, W. B., Fan, X. M., Huang, F. M., et al., 2021. Uncertainties of Landslide Susceptibility Modeling under Different Environmental Factor Connections and Prediction Models. _Earth Science_, 46(10): 3777-3795.

[15] Luo, L. G., Pei, X. J., Huang, R. Q., et al., 2021. Landslide Susceptibility Assessment in Jiuzhaigou Scenic Spot by CF and Logistic Regression Coupled with GIS. _Journal of Engineering Geology_, 29: 526-535.

[16] Yao, J., Qin, S., Qiao, S., et al., 2022. Application of a Two-Step Sampling Strategy Based on Deep Neural Network for Landslide Susceptibility Mapping. _Bulletin of Engineering Geology and the Environment_, 81: 1-20.

[17] Yao, X., Tham, L., Dai, F., 2008. Landslide Susceptibility Mapping Based on Support Vector Machine: A Case Study on Natural Slopes of Hong Kong, China. _Geomorphology_, 101: 572-582.

[18] Cina, A. E., 2022. A Black-Box Adversarial Attack for Poisoning Clustering. _Pattern Recognition_, 122: 108306.

[19] Semwal, P., 2022. Cyber-attack Detection in Cyber-Physical Systems Using Supervised Machine Learning. _Handbook of Big Data Analytics and Forensics_: 131-140.

[20] Engstrom, L., 2019. Exploring the Landscape of Spatial Robustness. _Proceedings of the 36th International Conference on Machine Learning_: 1802-1811.

[21] Szegedy, C., et al., 2014. Intriguing Properties of Neural Networks.

[22] Nguyen, A., et al., 2015. Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images. _Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition_: 427-436.

[23] Hengstler, M., 2016. Applied Artificial Intelligence and Trust - The Case of Autonomous Vehicles and Medical Assistance Devices. _Technological Forecasting and Social Change_, 105: 105-120.

[24] Cheng, K. Y., Wang, N., Shi, W. X., et al., 2022. Research Progress on the Interpretability of Deep Learning. _Computer Applications_, 42(12): 3639-3650.

[25] Adebayo, J., Gilmer, J., Muelly, M., et al., 2018. Sanity Checks for Saliency Maps. _Advances in Neural Information Processing Systems_.

[26] Yeh, C. K., Kim, B., Wexler, J., et al., 2019. On the (In) fidelity and Sensitivity of Explanations. _Advances in Neural Information Processing Systems_.

[27] Ghosal, S., Yoon, J., & Van der Schaar, M. (2022). Explaining by Removing: A Unified Framework for Model Explanation. arXiv preprint arXiv:2202.00872.

[28] Li, H., Wu, H., & Jin, R. (2022). Understanding Deep Neural Networks via Layer - wise Influence Estimation. In International Conference on Learning Representations.