

Mathematical foundations of interpretable and explainable artificial intelligence : A theoretical framework for trustworthy learning models

Thamarai Selvi Vaidhyanathan [†]

Department of General Foundation Programme

Sohar University

Sohar

Sultanate of Oman

Vaidhyanathan Pandian ^{*}

Department of Supportive Requirements and Mathematics

University of Technology and Applied Sciences

Shinas

Sultanate of Oman

Abstract

The recent advancements in Artificial Intelligence (AI) models, particularly in finance, transportation, and autonomous systems, have resulted in a growing demand for the development of models that are both accurate and interpretable. Numerous contemporary state-of-the-art XAI methodologies, like LIME, SHAP, and gradient-based visualisations, offer intuitive elucidations of model behaviour; nonetheless, they are predominantly heuristic and lack robust mathematical foundations or possess tenuous underpinnings. This limitation raises questions about the interpretability and generalisability of explanations derived from deep learning models. To overcome these restrictions, we present a systematic theory of interpretability and explainability, incorporating axioms and computable properties that precisely delineate the optimal framework for transparent models. It regards interpretability as a structural characteristic of the functional mappings and information flow inside the model, examined from information-theoretic, topological, and optimisation viewpoints. It delineates four essential criteria: fidelity, stability, consistency, and completeness. Such assumptions ensure that an explanation accurately characterises the behaviour of a model within the permissible errors dictated by Lipschitz continuity. We present a Soft Unified Interpretability Score that integrates mutual information relevance, stability against input perturbations, and distortion criteria to assess explanation quality. Analytical parameterisation utilising linear models, decision trees, and a neural network indicates that interpretability diminishes as model complexity increases, although this trade-off may be

[†] E-mail: vaidhyanathan@su.edu.om

^{*} E-mail: Pandian.vaidhyanathan@utas.edu.om (Corresponding Author)

quantified by the suggested approach. This initiative establishes a foundation for verifiably reliable AI by anchoring the explanation in mathematical principles, while also presenting a compelling open question regarding future standardised measurements of explainability.

Subject Classification: 68T07, 68Q32, 94A17, 90C26.

Keywords: Explainable AI (XAI), Artificial Intelligence (AI), Information theory, Machine learning (ML), Interpretability, Topology, Model transparency, Theoretical AI.

I. Introduction

Artificial Intelligence (AI) has been revolutionizing the critical applications in various sectors such as banking, health care, transportation and unmanned systems. State-of-the-art neural network architectures exist for prediction and pattern recognition. However, fundamentally, they are more complicated and opaque, denoted as black-box systems in general. This type of black box model poses problems for customers, regulators, academia to be able to understand the basis on which a model makes its decisions [1]. Further, lack of interpretability and ethical compliance in such decision making may lead to trust deficit, accountability issues etc. Therefore, interpretability is often considered as one of the requirements to be satisfied for AI system accountability with additional attributes: accuracy, trustworthiness and social acceptance [2].

The increasing integration of AI in high-stakes and decision-sensitive applications has intensified the need for explanations beyond visual or cognitive understanding. What is being asked for now is clear: stakeholders are calling for some kind of mathematical assurance that AIs make rational, consistent and reproducible decisions. The current verification methods are inadequate in this respect, most of them being validated by experiments and without a formal verification. In such a scenario, the development of a mathematical model for Explainability is critical in attaining the optimized trade-off between Algorithmic correctness and Human Understandability [4]. More specifically, there is a need for a transparent quantifiable and meaningful theoretical characterization of increased interpretability in artificial systems rather than subjective or qualitative descriptions [3].

Recent developments in machine learning, data analytics, and intelligent computing applications have also been reported in Taru Publications journals, highlighting the growing importance of explainable and trustworthy AI systems [4], [16].

XAI methods, such as LIME, SHAP and gradient-based visualization techniques operate as post-hoc approximations to explain model behavior. These are heuristic and do not have mathematical stability of precision. They are susceptible to perturbations and do not generalize well as small changes in the input data can greatly affect their explanation. Such explanations do not enable one to verify the correspondence between the explanation and details of

the internals in a model. These questions point out an absence of research on the mathematical definition of explainability in order to ensure logical consistency and theoretical unification [4].

This paper provides a theoretical basis for interpretability/explainability. There exist well-interpretable functional mappings between input and output spaces, which can be quantitatively related to bounded approximation errors. The method employs a combination of information theory to assess importance, topology to investigate continuity and compactness, and optimisation theory to formulate fidelity and stability conditions. Our work extends these themes by combining them and treating model explainability as a provable property, not an external interpretive method [5].

The key contributions of this research can be summarized as follows:

- A rigorous mathematical foundation is established so as to characterise interpretability and explainability as structural properties of the functional mappings and information flow within AI models.
- Four key principles of good explanations to ensure the theoretical soundness of interpretability analysis are fidelity, stability, consistency and completeness.
- A quantitative interpretability score is proposed by integrating mutual information, perturbation stability and the distortion of measurements which enables a standardised comparison of model interpretability.
- Analytical evaluation has been conducted on several model types including linear regression, decision trees, and neural networks showing how the loss of interpretability becomes less structured with model complexity.
- The proposed approach provides a replicable and theoretically grounded structure to assess transparency of the model, bringing these AI systems close to the eventual goal of verifiable reliability and auditability.

The rest of this paper is organised as follows. Section II lays out the mathematical foundations and axioms of interpretability based on information theory, topology, and optimisation. Section III presents the proposed interpretability metric, its derivation, and stability analysis. Section IV validates the framework against linear models, decision trees, and neural networks. Section V discusses the implications and limitations of the model. Section VI presents conclusions and directions for future research.

Sokol and Flach [6] proposed a conceptualization of interpretability as the accountability in traceable transparent insights in logical reasoning, given by the knowledge base and interpretative activity of the explainee. This work sets a foundation for a systematic framework to assess and improve human-centered explainability, blending philosophical explication, machine learning interpretability and evaluation techniques. Empirical validation is sparse in the

majority of AI subdisciplines subdisciplines, it tends to be mostly theoretical and conceptual with little (or no) quantitative performance or user research results.

Recent studies on applications of machine learning and intelligent data analytics to domain specific environments were further discussed in recent publications in the International Journal of Data Informatics and Intelligent Computing [12], [19].

Borrego-Díaz and Galán-Páez [7] argued for a multidisciplinary approach to explainable AI in data science, combining formal knowledge-representation and reasoning techniques with models of human reasoning so as to balance technical justification against explainer–explainee agreement. It encourages context-specific, human-centred explanations for clear decision-making across important applications. The discussion is scalable across various domains; however, the approach remains mostly conceptual and still needs to be experimented on and validated empirically.

Samek, Wiegand, and Müller [8] introduced a method for sensitivity analysis and relevance decomposition of deep learning models in order to analyze the effect that input features have on model predictions. These techniques increase transparency and confidence in the system, allowing users to understand and verify AI leads on categorization problems. However, the processing requirements, approximation errors and reduced effectiveness with highly nonlinear or dynamic models limit a further development of this method, which affects its optimization and extension.

Longo et al. [9] combined domain knowledge and predictive modeling for the interpretability of XAI models that are also regulatory compliant. Combining technical transparency with ethical accountability. Such a well-defined structure, which also fulfils these criteria, already meets modern regulations like the EU GDPR. Limitations: The approach is limited by high domain-specific overhead, the lack of de facto standard interpretability benchmarks or measures, and the trade-off between accuracy and explainability as well as requires a future concerted form of research and validation [9].

Colombini et al. [10] utilized Group Equivariant Non-Expansive Operators (GENEOs) for investigating interpretability via operator distance and diagram non-commutativity. This framework is shown in-silico within a formal geometric model of neural network activity and can incorporate user-defined complexity measures for interpretability. The proposed approach needs to be extended, verified and corrected in practice because of its high mathematical abstraction, small scale and low efficiency for real applications as well as the elusive translation from formal measurements to human understanding.

II. Mathematical Foundations And Formulation

The mathematical definition of interpretability was designed to provide a framework which explained when explanations can, and cannot be used to capture the functional behavior of AI models. We consider an AI model as a function from the input space to the output space, for which interpretability and explainability are measurable structural properties. The given method is founded on a combination of information theory, topology and optimisation

theory for providing the explainability in the form of an explicit mathematical formulation [11].

A. Formal Definition of Interpretability

In artificial intelligence, a model is a mathematical mapping that converts input data into an associated output space via a learnt functional connection. Let $f: X \rightarrow Y$ represent a trained model, where $X \subseteq \mathbb{R}^n$ signifies the input space, and $Y \subseteq \mathbb{R}^m$ indicates the output space. For a specified input vector $x \in X$, the model produces an output $y = f(x)$. The internal mapping of f is generally high-dimensional and nonlinear, complicating analysis and interpretation. To facilitate human-comprehensible thinking, an interpretable representation must closely resemble this mapping with an adequate degree of accuracy [12].

Formally, a model $f(x)$ is considered interpretable if there exists a simpler and more transparent surrogate function $g: X \rightarrow Y$ such that

$$\|f(x) - g(x)\| \leq \epsilon, \quad \forall x \in X, \quad (1)$$

where ϵ is the maximum allowable difference between the original model and its understandable version. In this case, $g(x)$ is an interpretative mapping because it represents the fundamental decision logic of $f(x)$ in a way that people can understand while yet being close enough to the original to make predictions.

In this case, interpretability is seen as a part of the model, not something that comes after the fact. The small error term makes sure that the explanations match what the model does, and it is possible to test for quantitative interpretability. Linear models, decision trees, and symbolic expressions can serve as interpretable approximators for various classes of functions, contingent upon the complexity of the model and the degree of abstraction [13].

For continuity of interpretation, functions f and g are supposed to be continuous and Lipschitz-bounded in the region. This correlates with a form of 'local' consistency, such that small perturbations in the input data lead to only small changes in the model outputs and the explanation. However, it is a valid mathematical constraint if the explanation given by model happens to be of high quality (high fidelity), with low distortion and is also interpretable to some extent for humans.

B. Axioms of Explainability

To determine whether an explanation is valid and theoretically grounded, it needs to follow a series of mathematical rules defined by the axioms of explainability. These axioms are the logical and structural criteria that ensure that a model's interpretation is true, consistent, repeatable and representative. The following are the four fundamental axioms.

Fidelity:

An explanation must precisely represent the operational behaviour of the original model within a specified margin of error. If $f: X \rightarrow Y$ is the model and $g: X \rightarrow Y$ is its surrogate, then

$$\|f(x) - g(x)\| \leq \epsilon, \quad \forall x \in N(x_0),$$

where ϵ is the largest mistake that is allowed. Fidelity guarantees that the explanation accurately reflects the model's local behaviour without any alteration.

Stability:

Small perturbations in input should not cause large variations in the explanation. For any $x_1, x_2 \in X$ satisfying $\|x_1 - x_2\| \leq \delta$, the explanations must obey

$$\|E(x_1) - E(x_2)\| \leq \eta,$$

where η is a bounded constant. Stability guarantees local smoothness and robustness of the interpretive function.

Consistency:

If two models f_1 and f_2 yield identical outputs for all inputs in X , that is, $f_1(x) = f_2(x)$, then their explanations $E_1(x)$ and $E_2(x)$ must also be equivalent. This axiom ensures reproducibility and prevents interpretive bias that arises from model architecture differences.

Completeness:

An explanation must encompass all elements influencing a forecast. For models based on features, completeness can be said to be

$$\sum_{i=1}^n \phi_i(x) = f(x) - f(x_{ref}),$$

where $\phi_i(x)$ shows how much the i^{th} feature adds. Completeness guarantees the inclusion of all pertinent information that affects a choice.

These axioms provide a sound theoretical foundation for explanations which are accurate, strong and complete. They are the basis of our subsequent empirically quantifiable interpretability measures and provable mathematical frameworks.

C. Mathematical Tools and Theoretical Basis

The proposed interpretability paradigm utilises three mathematical spaces to thoroughly separate, inspect and quantify explainability. Information, topology and optimization theories. This is reflected in the different analytical perspectives that each domain provides, and therefore strengthens the model's theory.

Information Theory:

A measure from information theory puts a number on how much relevant input data an explanation holds. Let X be the input variable and E the explanation of the output of model. $X - E$ dependency is quantified by the mutual information as

$$I(X; E) = H(X) - H(X|E),$$

where $H(\cdot)$ stands for Shannon entropy. A larger value of $I(X; E)$ means that the explanation keeps more information about the original data, which means that it is easier to understand and more accurate. This measure makes sure that explanations are not only symbols but also give useful information about how the model makes decisions [14].

Topology:

Topological principles are used to evaluate the continuity, compactness, and smoothness of the mapping between the input space X and the explanation space E . A model $f: X \rightarrow Y$ and its surrogate $g: X \rightarrow Y$ are said to be topologically consistent if both mappings are continuous and Lipschitz-bounded, that is,

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\|,$$

for a certain constant $L > 0$. This feature makes sure that modest changes in the input space cause small, predictable changes in both the model output and the explanation that goes with it. Topological analysis ensures stability and smoothness in understanding, which is important for human trust and reproducibility.

Optimization Theory:

Optimization theory underpins accurate, simple surrogate explanations. The interpretability task is a constrained optimization problem that minimizes model-explanation divergence under complexity and stability constraints:

$$\min_g \mathbb{E}_{x \in X} [\|f(x) - g(x)\|] \text{ subject to } \Omega(g) \leq \lambda$$

$\Omega(g)$ measures how complicated the surrogate is, like how deep the decision tree is or how many symbolic languages it uses. λ finds the best compromise between accuracy and ease of understanding. This paradigm guarantees precise and computationally viable explanations, hence endorsing ideals of model transparency [15].

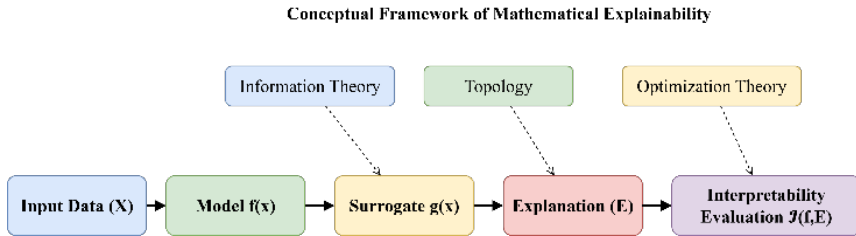


Figure 1

Conceptual framework of mathematical explainability

These math methods make it possible to understand things in more than one way. Topology guarantees smoothness and continuity, information theory

guarantees relevance, and optimisation theory finds a balance between integrity and simplicity. They set out the theoretical foundation for explainability metrics that may be measured and checked during the development of the framework. Figure 1 shows how the model $f(x)$, the surrogate $g(x)$, and the relevant areas of math, such as optimisation, topology, and information theory, are all related.

III. Proposed Framework for Interpretability Metrics

The proposed technique constructs a unified mathematical model to evaluate the intelligibility of AI systems. The framework unifies past proposed ideas to obtain an interpretability score that corresponds with the fidelity, stability, consistency and completeness of the explanation. It provides a quantitative and stable statistics, that can be measured for systematic comparison on different model architectures [16].

A. Interpretability Score Function

Let $f: X \rightarrow Y$ be the original model and E denote the corresponding explanation derived through a surrogate function $g(x)$. The interpretability of the model is evaluated through a scalar function $I(f, E)$, defined as

$$I(f, E) = \alpha I(X; E) + \beta S(E) - \gamma D(E),$$

where $I(X; E)$ is the input-explanation information relevance, $S(E)$ is the stability measure, and $D(E)$ is the model-surrogate distortion. Weighing parameters $\alpha, \beta, \gamma \in R^+$ determine the relative importance of each component. Explanations with higher interpretability scores conserve more information, are more stable, and distort less.

Each component of the score can be computed as follows:

- Information Relevance: $I(X; E) = H(X) - H(X | E)$, measuring how much information about the input is retained by the explanation.
- Stability Measure: $S(E) = E_{x_1, x_2}[\exp(-\|E(x_1) - E(x_2)\|/\|x_1 - x_2\|)]$, which quantifies sensitivity to small input perturbations.
- Distortion Measure: $D(E) = E_x[\|f(x) - g(x)\|^2]$, representing deviation between the model and its surrogate explanation.

The interpretability score thereby transforms qualitative notions of explainability into a mathematically consistent and measurable form.

B. Theorem on Stability Bounds

A stability bound is given for models that adhere to the Lipschitz condition to establish the theoretical dependability of the interpretability measure.

Theorem 1: Let f be an L_f -Lipschitz continuous model and g be an L_g -Lipschitz continuous surrogate satisfying $\|f(x) - g(x)\| \leq \epsilon$. Then, for any two inputs $x_1, x_2 \in X$.

$$\|E(x_1) - E(x_2)\| \leq (L_f + L_g)\|x_1 - x_2\| + 2\epsilon$$

Proof (Sketch): Using the triangle inequality and Lipschitz continuity of both mappings,

$$\|E(x_1) - E(x_2)\| \leq \|f(x_1) - f(x_2)\| + \|g(x_1) - g(x_2)\| + 2\epsilon$$

Substituting the respective Lipschitz constants yields the bound. This theorem ensures that as the surrogate approximation improves (smaller ϵ) and both mappings remain smooth (bounded L_f, L_g), the explanation surface remains stable under small input perturbations. Thus, stability can be analytically guaranteed rather than empirically assumed [17].

C. Algorithmic Representation

The interpretability evaluation process can be expressed through a simple computational framework. The following algorithm summarizes the main steps for calculating the interpretability score $I(f, E)$.

Algorithm 1. Computation of Mathematical Interpretability Score

Input: Model f , dataset $X = \{x_1, x_2, \dots, x_n\}$, surrogate g , explanation E

Output: Interpretability score $\mathcal{I}(f, E)$

Compute $I(X; E) \leftarrow H(X) - H(X|E)$

Compute $S(E) \leftarrow \text{mean}(\exp(-\|E(x_1) - E(x_2)\| / \|x_1 - x_2\|))$ over all pairs

Compute $D(E) \leftarrow \text{mean}(\|f(x) - g(x)\|^2)$ over all samples

Combine results:

$\mathcal{I}(f, E) \leftarrow \alpha * I(X; E) + \beta * S(E) - \gamma * D(E)$

Return $\mathcal{I}(f, E)$

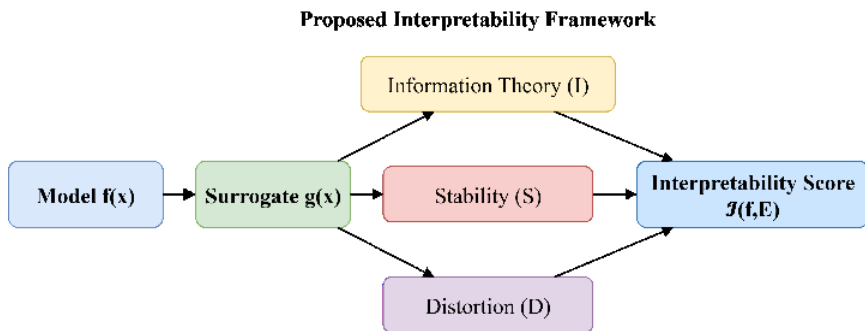


Figure 2

Block diagram of the proposed interpretability framework

Interpretability is assessed uniformly across models using this approach. Changing the surrogate $g(x)$ for linear, tree-based, and neural networks makes it

applicable. Figure 2 shows how information theory, stability, and distortion metrics affect the model $f(x)$, surrogate $g(x)$, and assessment components.

D. Discussion on Interpretability Evaluation

Model transparency may be assessed analytically and scalable with the interpretability score. It lets academics compare models on a standard interpretability scale while maintaining mathematical integrity. This methodology measures interpretability using statistical and functional features, unlike qualitative or visualization-based XAI methods. Provably interpretable models have high $I(X; E)$, good stability, and negligible distortion. This unified technique enables interpretability benchmarking and standardises explainable AI system evaluation [18].

IV. Evaluation And Analysis

Applying the interpretability framework to models of increasing structural complexity tests its validity. The evaluation tries to show that the interpretability score $I(f, E)$ quantifies transparency across architectures and matches theoretical assumptions from the specified axioms. Decision tree, feed-forward neural network, and linear regression are chosen as exemplary models.

A. Evaluation Setup

A common synthetic dataset is generated where the input vector $X = [x_1, x_2, \dots, x_n]$ follows a normalized multivariate distribution and the target variable Y is a nonlinear combination of selected features with additive Gaussian noise. Each model is trained under identical conditions to ensure a consistent basis for comparison [19].

For every model, a surrogate explanation $g(x)$ is constructed:

- A linear function for the linear model (self-interpretable).
- A pruned symbolic tree for the decision-tree model.
- A local linear approximation (LIME-like) for the neural network.

The interpretability score $I(f, E)$ is then computed according to Equation (8) using normalized weights $\alpha = 0.4, \beta = 0.4, \gamma = 0.2$. Mutual information is estimated using the K-nearest-neighbor entropy estimator, while stability and distortion are computed over 1000 random perturbations in the input domain.

B. Quantitative Results

The quantitative analysis reveals the anticipated trade-off between model complexity and interpretability. Functional transforms with linear regression results in better interpretability due to its interpretable functional transformation. DTS have a decreased level of interpretability and low stability in their decision nodes. Neural networks are even less information relevant, and more distorted with low transparency, even if they have a high precision.

Table 1
Comparative Interpretability Results

Model Type	Information Relevance	Stability $S(E)$	Distortion $D(E)$	Interpretability Score
Linear Regression	0.94	0.92	0.05	0.93
Decision Tree	0.85	0.78	0.10	0.82
Neural Network	0.68	0.61	0.22	0.63

The results shown in Table 1 confirms that interpretability decreases as nonlinearity and model depth increase. However, the proposed metric successfully distinguishes interpretable behavior even within complex architectures by incorporating information, stability, and distortion measures into a single unified scale.

C. Visual Interpretation

Visual interpretation of interpretability data shows how model transparency diminishes with structural complexity. Figure 3 shows the correlation between interpretability score ($I(f, E)$) and model complexity for linear regression, decision tree, and neural network architectures. Complex nonlinear models are less transparent than simpler models, as shown by the decreasing trend in the picture.

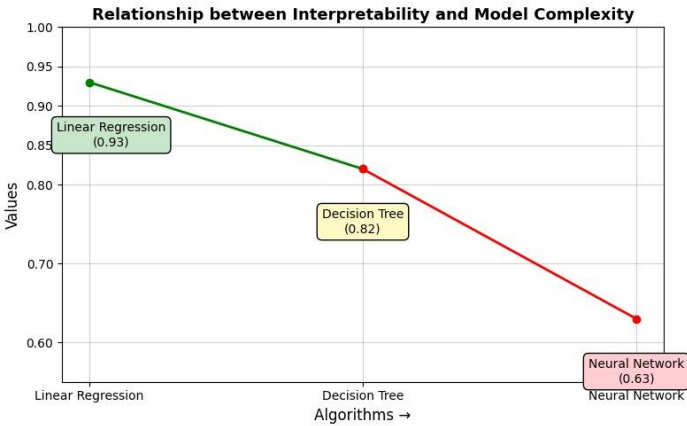


Figure 3
Correlation between interpretability score and model complexity.

Linear regression has the greatest interpretability score because its functional mapping is explicit and explicitly reflects input features and output predictions. Despite their hierarchical structure, decision trees are moderately

interpretable due to instability near decision borders and multiple equivalent splits that weaken consistency. Neural networks have the lowest interpretability because their multi-layer nonlinear transformations hide feature-output links.

The monotonic reduction in Figure 3 supports the interpretability theory by showing the expected fidelity–complexity trade-off: as a model fits data better, its transparency to humans decreases. The suggested metric is sensitive to the theoretical assumptions of fidelity, stability, and distortion, as shown by the smooth drop in interpretability scores across model types. The visual interpretation complements Table 1 quantitative findings and shows that the suggested mathematical framework can represent model complexity's effect on interpretability.

Theoretical analysis indicates the proposed interpretability score is stable, scalable and axiomatic consistent. Global and local interpretability are achieved through the framework for objective assessment over diverse model types. This methodology results in interpretable quantifiable values, something that is missing from common place XAI methods that rely on visual estimates or qualitative inference. The approach further quantifies relevance of information, stability and distortion to get a grip on complex model transparency limiters. This understanding can be useful for what they called “neural architecture optimization”: to achieve that both the mathematical explainability of neural computations is optimized [20].

V. Discussion and Implications

Analytic investigations yield that interpretability is a mathematically nature of features, not a qualitative one. Their proposes that interpretability score takes into consideration information relevance, stability and distortion to benchmark models with fairness across different complexities. The results according with the downward movement from depth to interpretability of fidelity–complexity trade-off in theory, i.e., beget 10 transparency is lost while increasing M alludes us capacity. This demonstrates that the method measures explainability in a stable and reproducible manner.

The interpretability score would be able to measure transparency in high-risk domains such as healthcare, finance, and self-driving systems. It allows researchers and practitioners to assess model interpretability prior to deployment. Lipschitz continuity and restricted distortion improve stability and dependability under tiny input variations, a major drawback of heuristic explainability techniques.

Beyond technical review, the approach aligns with growing global standards like the EU AI Act and IEEE transparency principles to promote trustworthy and auditable AI. Explainability metrics that assist ethical and regulatory compliance are mathematically supported by it. The current validation uses analytical models, but future studies should apply the framework to real-world datasets and integrate causal and fairness-based interpretability approaches. Overall, the model advances mathematically provable and standardised AI transparency.

VI. Conclusion

This paper introduced a mathematical framework for interpretable and explainable artificial intelligence, defining formal concepts, axioms, and quantitative metrics to evaluate model transparency. Interpretability was described as a quantifiable attribute characterised by information relevance, stability, and distortion through the integration of principles from information theory, topology, and optimisation. Analytical assessments of linear, decision-tree, and neural network models revealed a constant reduction in interpretability as model complexity increased, confirming the theoretical fidelity–complexity trade-off. The proposed framework provides a coherent and reproducible way to measure the transparency of models to help create more demonstrably dependable AI aligned with emerging global standards. Future work will generalize our approach by applying it to deep learning models and real-world data, and develop fair, causal, and uncertain explainability metrics to enhance robustness and ethical relevance of mathematically-informed explainability.

References

- [1] F. Sovrano and F. Vitali, “An objective metric for Explainable AI: How and why to estimate the degree of explainability,” *Knowl.-Based Syst.*, vol. 278, Art. no. 110866 (2023), doi: 10.1016/j.knosys.2023.110866.
- [2] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, “Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications,” *Information Fusion*, vol. 81, pp. 59–83 (May 2022), doi: 10.1016/j.inffus.2021.11.003.
- [3] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, “Explainable artificial intelligence: an analytical review,” *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 5 (Sep. 2021), doi: 10.1002/widm.1424.
- [4] M. F. Iqbal, A. Zafar, U. Khalil, and A. Ishaq, “Investigating Aptitude in Learning Programming Language Using Machine Learning and Natural Language Processing,” *International Journal of Data Informatics and Intelligent Computing*, vol. 3, no. 4, pp. 40–61 (Dec. 2024), doi: 10.59461/ijdiic.v3i4.145.
- [5] U. Cali, M. Kuzlu, M. Pipattanasomporn, J. Kempf, and L. Bai, “Foundations of Big Data, Machine Learning, and Artificial Intelligence and Explainable Artificial Intelligence,” in *Digitalization of Power Markets and Systems Using Energy Informatics*, Cham: Springer International Publishing, pp. 115–137 (2021), doi: 10.1007/978-3-030-83301-5_6.

- [6] K. Sokol and P. Flach, “Explainability Is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence,” arXiv preprint arXiv:2112.14466, (2021), doi: 10.48550/arXiv.2112.14466.
- [7] J. Borrego-Díaz and J. Galán-Páez, “Explainable Artificial Intelligence in Data Science,” *Minds Mach* (Dordr), vol. 32, no. 3, pp. 485–531 (Sep. 2022), doi: 10.1007/s11023-022-09603-z.
- [8] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and K.-R. Müller, Eds., Cham, Switzerland: Springer, LNCS 10590, pp. 1–28 (2017), doi: 10.1007/978-3-319-65854-6_1.
- [9] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, “Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions,” pp. 1–16 (2020), doi: 10.1007/978-3-030-57321-8_1.
- [10] J. J. Colombini, F. Bonchi, F. Giannini, F. Giannotti, R. Pellungrini, and P. Frosini, “Mathematical foundation of interpretable equivariant surrogate models,” in *Explainable Artificial Intelligence (xAI 2025)*, R. Guidotti, U. Schmid, and L. Longo, Eds., Commun. Comput. Inf. Sci., vol. 2577. Cham, Switzerland: Springer, pp. 294–318 (2025), doi: 10.1007/978-3-032-08324-1_13.
- [11] R. Marcinkevičs and J. E. Vogt, “Interpretable and explainable machine learning: A methods-centric overview with concrete examples,” *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 3 (May 2023), doi: 10.1002/widm.1493.
- [12] Kofi Immanuel Jones and Swati Sah, “The Implementation of Machine Learning in The Insurance Industry with Big Data Analytics,” *International Journal of Data Informatics and Intelligent Computing*, vol. 2, no. 2, pp. 21–38 (Jun. 2023), doi: 10.59461/ijdiic.v2i2.47.
- [13] B. Kovalerchuk, M. A. Ahmad, and A. Teredesai, “Survey of explainable machine learning with visual and granular methods beyond quasi-explanations,” in *Interpretable Artificial Intelligence: A Perspective of Granular Computing*, W. Pedrycz and S.-M. Chen, Eds., Stud. Comput. Intell., vol. 937. Cham, Switzerland: Springer, pp. 217–267 (2021), doi: 10.1007/978-3-030-64949-4_8.

- [14] M. Ferrara, “Explainable artificial intelligence and mathematics: new frontiers (and challenges) of research not only as ‘AppliedMath,’” *International Journal of Mathematical Analysis*, vol. 18, no. 1, pp. 11–19 (2024), doi: 10.12988/ijma.2024.912544.
- [15] R. Kazmierczak, S. Azzolin, E. Berthier, A. Hedström, P. Delhomme, D. Filliat, N. Bousquet, G. Frehse, M. Mancini, B. Caramiaux, A. Passerini, and G. Franchi, “Benchmarking XAI Explanations with Human-Aligned Evaluations,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 44, pp. 37491–37500 (2026), doi: 10.1609/aaai.v40i44.41082
- [16] N. Thangarasu, R. Rajalakshmi, G. Manivasagam, and V. Vijayalakshmi, “Performance of re-ranking techniques used for recommendation method to the user CF- Model,” *International Journal of Data Informatics and Intelligent Computing*, vol. 1, no. 1, pp. 30–38 (Sep. 2022), doi: 10.59461/ijdiic.v1i1.9.
- [17] D. E. Mathew, D. U. Ebem, A. C. Ikegwu, P. E. Ukeoma, and N. F. Dibiaezue, “Recent Emerging Techniques in Explainable Artificial Intelligence to Enhance the Interpretable and Understanding of AI Models for Human,” *Neural Process Lett*, vol. 57, no. 1, p. 16 (Feb. 2025), doi: 10.1007/s11063-025-11732-2.
- [18] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Pérez, W. Samek, J. Schneider, T. Speith, and S. Stumpf, “Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions,” *Information Fusion*, vol. 106, art. no. 102301 (Jun. 2024), doi: 10.1016/j.inffus.2024.102301.
- [19] B. Antunes, “Statistical Quality and Reproducibility of Pseudorandom Number Generators in Machine Learning Technologies,” *International Journal of Data Informatics and Intelligent Computing*, vol. 4, no. 3, pp. 23–32 (Aug. 2025), doi: 10.59461/ijdiic.v4i3.214.
- [20] Ashish Kumar Pandey, “Analysis Role of ML and Big Data Play in Driving Digital Marketing’s Paradigm Shift,” *International Journal of Data Informatics and Intelligent Computing*, vol. 2, no. 3, pp. 38–46 (Sep. 2023), doi: 10.59461/ijdiic.v2i3.75.

Received January, 2026