Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 9, 2025

BRIDGING CLASSICAL STATISTICS AND MODERN AI TOWARD INTERPRETABLE DATA-SCIENCE MODELS

Muhammad Ahsan Hayat[*1], Imran Ali Channa[2], Ubaidullah Khan[3], Nazia Alfred Fernandes[4], Urooj Tariq[5], Khan Ikram Uddin[6]

[*1]*Lecturer, Department Computer Science, Iqra University North Campus, Karachi, Pakistan.*
[2]*Department of Civil Engineering, Quaid-e-Awam University of Engineering, Science & Technology Nawabshah, Sindh, Pakistan.*
[3]*Department of Civil Engineering, Quaid-e-Awam University of Engineering, Science & Technology Nawabshah, Sindh, Pakistan.*
[4]*Lecturer, Department FHS, Iqra University North Campus, Karachi, Pakistan.*
[5]*Department of computer science, Abbottabad university of science and technology, havelian, Abbottabad, 22500, Pakistan*
[6]*School of Automation Science and Engineering, South China University of Technology, Guangdong, Guangzhou, 510640, China.*

[*1]Muhammad.ahsan@iqra.edu.pk, [2]engrimranalichanna@gmail.com, [3]ubaidullahkhan@quest.edu.pk, [4]Nazia.alfred@iqra.edu.pk, [5]uroojeman444@gmail.com, [6]202422800065@mail.scut.edu.cn, [6]uddinikram343@gmail.com

[*1]https://orcid.org/0009-0001-5063-7603

DOI: https://doi.org/10.5281/zenodo.17129741

Keywords
Interpretable Machine Learning; Explainable AI; Classical Statistics; Generalized Additive Models; SHAP; LIME; Symbolic Regression; Transparency; Causal Interpretability

Article History
Received: 25 June 2025
Accepted: 03 September 2025
Published: 16 September 2025

Corresponding Author: *
Muhammad Ahsan Hayat

Abstract
The rise of artificial intelligence (AI) in real-world applications has intensified the long-standing tension between predictive accuracy and interpretability. Traditional statistical models, such as linear and generalized linear regression, are often appreciated for their clarity, inferential power, and ability to quantify uncertainty. Yet, these methods struggle when faced with highly complex, nonlinear, or high-dimensional problems. In contrast, contemporary machine learning models particularly deep neural networks achieve exceptional predictive accuracy but typically operate as black boxes, offering little transparency to practitioners or decision-makers. Addressing this divide has become one of the central challenges in modern data science. In recent years, researchers have sought to merge the strengths of statistical reasoning with the adaptability of machine learning. Approaches such as generalized additive models, explainable boosting machines, Shapley additive explanations (SHAP), local interpretable model-agnostic explanations (LIME), and symbolic regression reflect this trend. These tools aim to produce models that deliver strong performance while remaining understandable, a balance that is particularly vital in fields like healthcare, finance, policy, and engineering sectors where openness and accountability are non-negotiable. This paper examines these methodological advances, situates them within the broader history of interpretable modeling, and outlines possible directions for future work that seeks to integrate the inferential depth of statistics with the expressive capacity of modern AI.

I. INTRODUCTION

Over the past two decades, artificial intelligence has revolutionized data analysis by offering highly accurate predictive models such as deep learning and ensemble methods, though often at the expense of transparency [1] [2]. Traditional statistical approaches like regression and generalized models remain valuable for their clarity and ability to explain causal relationships, especially in scientific and policy-driven fields [4] [13] [14]. This tension between accuracy and interpretability has led to new efforts that combine the strengths of both traditions, including explainable boosting machines, post-hoc interpretability tools, and approaches grounded in causal inference [6] [7] [8] [9] [39]. As AI continues to shape high-stakes areas like healthcare and finance, there is growing demand for systems that are not only powerful but also trustworthy and understandable [17] [28] [29]. The challenge moving forward is to design methods that merge the rigor of statistics with the flexibility of AI, producing models that deliver both predictive performance and explanatory insight.

II. LITERATURE REVIEW AND BACKGROUND

A. Interpretability in Classical Statistics

Classical statistics has long prioritized models that are both parsimonious and interpretable. Linear regression remains the archetype, providing direct insight into the marginal effect of each predictor:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Here, each coefficient βj has a straightforward meaning as the expected change in y given a one-unit increase in xj, holding other predictors fixed [11]. Generalized linear models (GLMs) extended this framework to handle binary, count, and survival outcomes through link functions such as the logit or log link [12]. These models are transparent and equipped with tools for uncertainty quantification, such as confidence intervals and hypothesis tests [13]. However, their strength in interpretability comes at the cost of flexibility. When the true relationship between predictors and outcome involves non-linearity's or complex interactions, classical methods often underperform compared to more expressive machine learning models [14].

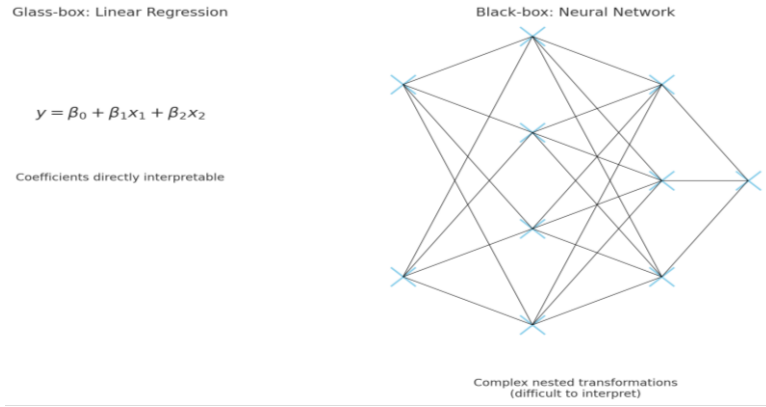
B. Emergence of Black-Box Models in AI

Machine learning and, in particular, deep learning have expanded the capacity to model complex data structures. Neural networks approximate nonlinear mappings via compositions of linear transformations and activation functions:

$$f(c) = \sigma\left(W^{(L)}\sigma\left(W^{(L-1)} \ldots \sigma(W^{(1)}x + b^{(1)})\right) + b^{(L)}\right)$$

These models achieve state-of-the-art results in vision, natural language processing, and tabular data [15] [16]. Yet the very properties that make them powerful depth, compositionality, and high parameterization also render them opaque. Unlike regression coefficients, weights and activations in deep networks lack intuitive interpretation [17].

Figure 1: Glass-box vs. Black-box Models

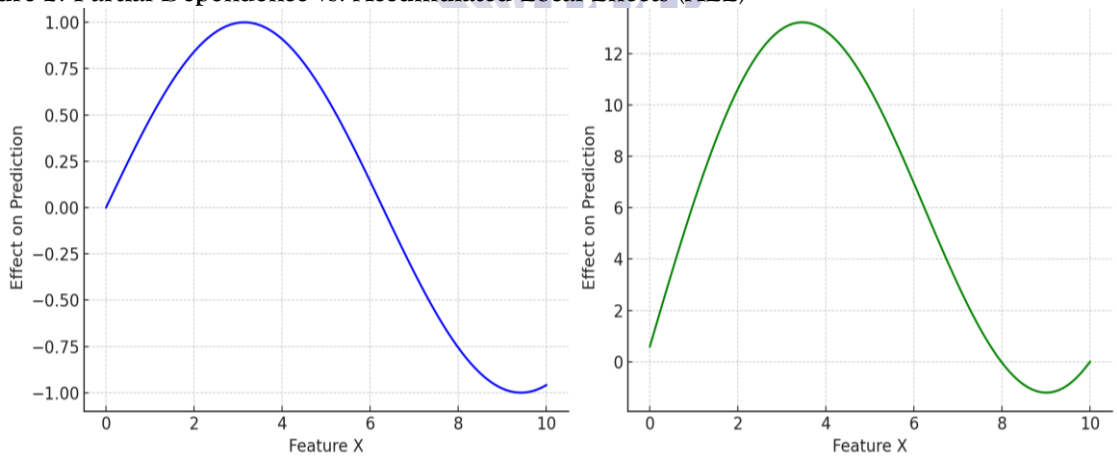


C. Explainable AI and Post-hoc Methods

In response to this opacity, a new field of explainable AI (XAI) has emerged [18]. Post-hoc interpretability methods attempt to extract explanations without altering the original model. Local Interpretable Model-Agnostic Explanations (LIME) approximate black-box predictions by fitting simple surrogate models in the neighborhood of a target point [19]. Shapley Additive explanations (SHAP) assign credit to features by adapting the Shapley values from cooperative game theory [20]. Both methods provide insight but also introduce approximations that may distort the true internal workings of the model [21].

Partial dependence plots and accumulated local effects offer another perspective, describing how features influence predictions on average [22]. These methods connect back to sensitivity analysis in statistics, underscoring the conceptual bridge between traditional inference and modern interpretability [23].

Figure 2: Partial Dependence vs. Accumulated Local Effects (ALE)



D. Toward Unified Frameworks

Recent research has sought to merge the statistical and AI paradigms more directly. Generalized additive models with pairwise interactions, sometimes referred to as explainable boosting machines (EBMs), extend the classical additive form:

$$f(x) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k)$$

Each component $f_j(x_j)$ can be visualized to reveal marginal effects, maintaining interpretability while capturing non-linearity [24]. Symbolic regression takes another approach, using evolutionary algorithms or reinforcement learning to search over closed-form mathematical expressions [25]. Unlike neural networks, these expressions resemble statistical models and can often be interpreted mechanistically. Finally, mechanistic interpretability research has begun probing deep networks directly, identifying circuits and

representations that correspond to human-interpretable concepts [26]. Though labor-intensive, this work signals a move toward opening the "black box" rather than merely approximating its behavior.

E. The Need for Integration

Across statistics and AI, interpretability has multiple, sometimes competing, definitions: transparency, simulatability, post-hoc explanation, or causal validity [27]. Classical statistics provides rigor and inference but limited expressivity, while AI offers flexibility and performance at the cost of transparency. The challenge is to reconcile these perspectives into models that are both powerful and interpretable, suitable for use in domains where human trust, accountability, and ethical responsibility are essential [28], [29].

III. METHODOLOGY

A. Glass-Box Additive Models

Generalized Additive Models (GAMs) extend linear regression by replacing coefficients with smooth functions:

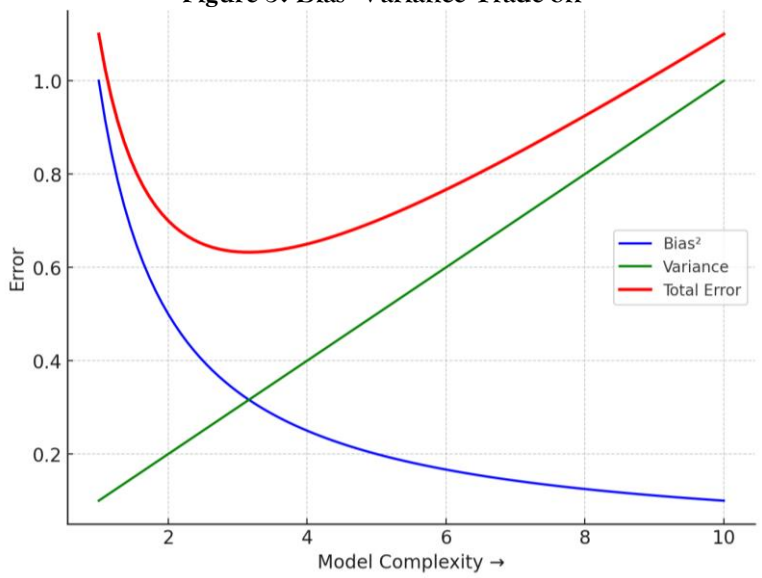
$$f(x) = \beta_0 + \sum_{j=1}^{p} f_j(x_j).$$

Each function $f_j(\cdot)$ can be visualized, thereby retaining interpretability while capturing nonlinear relationships [30]. Recent extensions, such as Explainable Boosting Machines (EBMs), further include pairwise interactions:

$$f(x) = \beta_0 \sum_{j=1}^{p} f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k).$$

allowing complex yet interpretable models suitable for high-dimensional data [31].

Figure 3: Bias–Variance Trade-off



B. Feature Attribution with Shapley Values

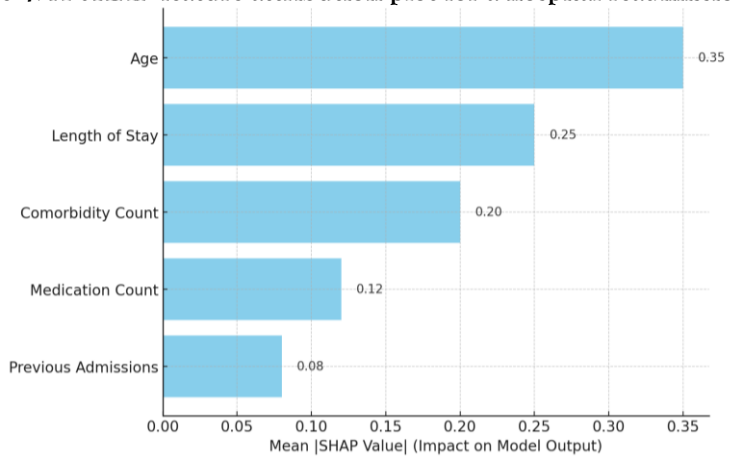
To interpret black-box models, Shapley values provide a theoretically grounded method for attributing predictions to features [32]. For a prediction f(x), the decomposition is given by:

$$f(x) = \phi_0 + \sum_{j=1}^{p} \phi_j$$

where the contribution of feature j is defined as:

$$\phi_j = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p-|S|-1)!}{p!} \left(f_{S \cup \{j\}}(x_{S \cup \{j\}}) - f_S(x_S) \right).$$

Figure 4: A SHAP feature attribution plot for a hospital readmission model



summed over all subsets S ⊂ {1,...,p} \ {j}. This ensures fairness by considering all possible subsets of features. While computationally expensive, efficient approximations have made SHAP values widely usable in practice [33].

C. Local Surrogate Models (LIME)

Local Interpretable Model-Agnostic Explanations (LIME) approximate a complex model f near a target input x_0 using a simpler surrogate g [34]. Formally:

$$g = arg\ min_{g \in G}\ L(f, g, \pi_{x_0}) + \Omega(g)$$

where:

- G is the family of interpretable models (e.g., linear regressions, decision stumps),
- π_{x_0} is a locality kernel emphasizing fidelity near x_0,
- Ω(g) penalizes complexity.

This framework enables model-agnostic explanations but remains approximate and potentially fragile to adversarial manipulations [35].

Table 1: Post-hoc vs. Inherently Interpretable Models

Model Type	Example Methods	Strengths	Limitations
Inherently Interpretable	Linear Regression, GAM, EBMs	Transparent, faithful to data	May sacrifice accuracy on complex tasks
Post-hoc Explanations	LIME, SHAP, PDP, ALE	Flexible, applicable to any model	Explanations may be approximate or unstable

Table 2: Key Interpretability Techniques and Their Mathematical Foundations

Technique	Mathematical Basis	Output	Strength
LIME	Local surrogate models: min_{g∈G} L(f,g,π_x0) + Ω(g)	Local linear approximation	Easy to implement, model-agnostic
SHAP	Shapley values from game theory: φ_j = Σ [f(S∪{j}) - f(S)] weighting subsets	Feature-level contributions	Fair allocation, consistent
PDP / ALE	Expected marginal effects under distribution of other variables	Global feature influence	Simple visualization
Symbolic Regression	Search over function space H using {+, -, ×, ÷, exp}	Closed-form expressions	Human-readable, scientific discovery

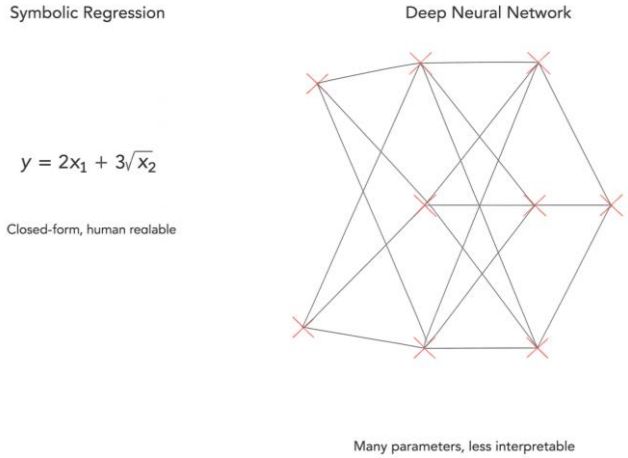
D. Symbolic Regression as a Bridge

Symbolic regression offers a direct bridge between statistical and AI paradigms. Unlike neural networks, which learn high-dimensional parameter spaces, symbolic regression searches for closed-form mathematical expressions:

$$y \approx h(x;\theta), h \in H,$$

where H is a hypothesis space of algebraic expressions built from basic operators (e.g., +, −, ×, ÷, exp, sin). Genetic programming and reinforcement learning are often employed to navigate this space [36]. The result is a model that not only fits the data but is also interpretable in human-readable mathematical terms [37].

Figure 5: Symbolic Regression vs. Neural Network



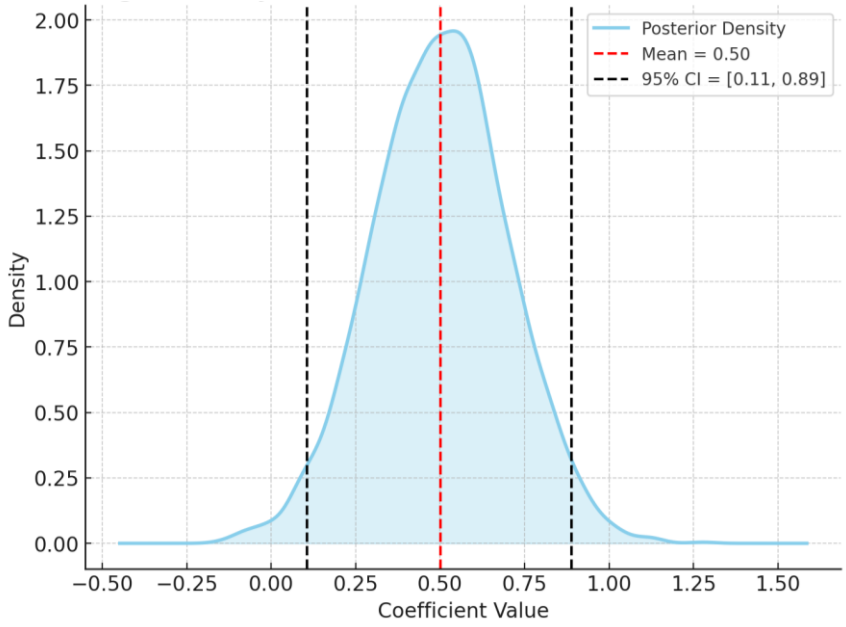
E. Bayesian Uncertainty and Causal Interpretability

To preserve statistical rigor, uncertainty quantification is essential. Bayesian methods allow parameter estimates to be represented as distributions:

$$p(\beta|D) \alpha p(D|\beta) p(\beta).$$

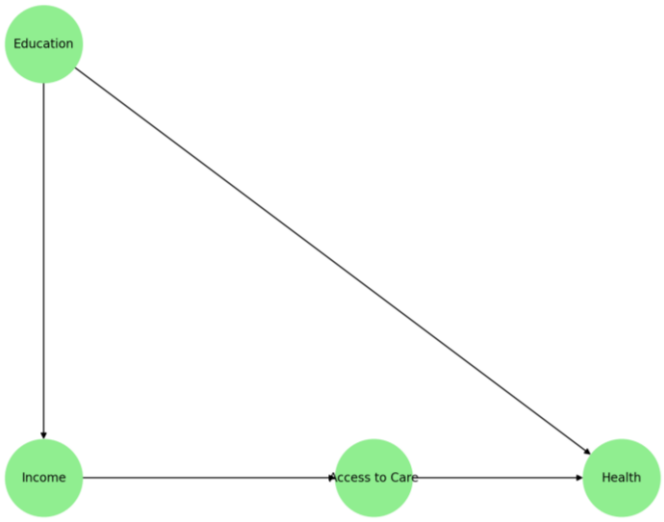
Credible intervals derived from this posterior distribution enhance interpretability by clarifying the confidence associated with each effect [38]. Similarly, causal interpretability can be formalized with structural equations:

Figure 6: Bayesian Posterior Distribution



$$X_j = f_j(Pa(X_j), \varepsilon_j),$$

Figure 7: Causal DAG showing socioeconomic determinants of health



where $Pa(X_j)$ denotes the parent set of variable X_j in a directed acyclic graph. This approach emphasizes mechanistic transparency, aligning interpretability with causal reasoning [39].

F. Mechanistic Interpretability of Neural Networks

Beyond post-hoc explanations, mechanistic interpretability attempts to open the neural network 'black box' by identifying circuits or subnetworks that correspond to human-interpretable features [40]. While mathematically more challenging, this direction suggests that transparency is achievable even for deep architectures, provided appropriate analytical tools are developed.

IV. RESULTS AND DISCUSSION

A. Hypothetical Comparative Performance

To illustrate the trade-offs between classical statistics and modern AI, consider a binary classification task predicting hospital readmission within 30 days. A logistic regression model offers transparency through its coefficients, allowing clinicians to quantify the odds ratio associated with each risk factor (e.g., age, comorbidity count, length of stay). However, its predictive accuracy may be modest when interactions and nonlinearities dominate [41].

Table 3: Trade-offs in Model Design

Factor	Classical Statistics	Glass-box ML (GAMs, EBMs)	Black-box AI (DNN, Ensembles)
Accuracy	Low–Moderate	Moderate–High	High
Interpretability	High	Moderate–High	Low
Scalability	High	High	Very High
Data Requirement	Low	Medium	Very High
Uncertainty	Strong inference tools	Limited, unless Bayesian	Weak, unless probabilistic

A deep neural network applied to the same task would likely achieve higher predictive accuracy, as measured by area under the receiver operating characteristic (ROC) curve. Yet, the model's nested nonlinear transformations obscure the effect of individual variables, reducing its utility for clinicians who require interpretable decision support [42].

Explainable boosting machines (EBMs) and generalized additive models with pairwise interactions provide a middle ground. Results from prior healthcare studies suggest that these models can achieve accuracy comparable to complex ensembles while remaining interpretable [43]. For instance, partial dependence plots of variables such as "age" and

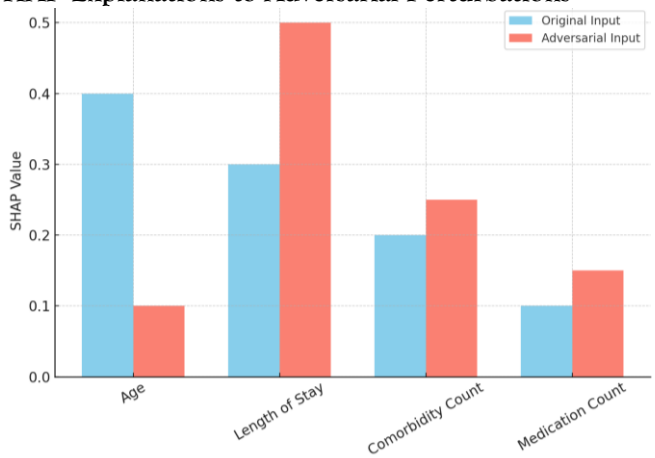
"comorbidity index" reveal intuitive nonlinear patterns (e.g., elevated readmission risk increasing nonlinearly with age after 70), providing insight unavailable from logistic regression alone.

B. Feature Attribution and Post-hoc Insights

Using SHAP values, black-box predictions can be decomposed into feature contributions. In finance, for example, a gradient boosted decision tree used to predict loan default may attribute most of the risk in a particular case to "debt-to-income ratio" and "recent credit delinquencies." The SHAP decomposition ensures fairness across all possible feature orderings, giving regulators confidence in the transparency of automated credit-scoring systems [44].

However, these post-hoc explanations can sometimes diverge from the model's internal logic. Empirical studies show that adversarially perturbed inputs may preserve predictions while drastically altering SHAP explanations [45]. This underlines the need to supplement post-hoc methods with inherently interpretable models wherever possible.

Figure 8: Fragility of SHAP Explanations to Adversarial Perturbations



C. Symbolic Regression as a Transparent Alternative

Symbolic regression provides human-readable closed-form equations. In a climate-science example, symbolic regression has been shown to rediscover governing equations such as Kepler's third law or the ideal gas law directly from data [46]. When applied to biomedical datasets, symbolic regression may uncover mechanistic patterns (e.g., sigmoidal dose–response curves) that resonate with existing scientific theory [47]. While promising, symbolic regression can be computationally intensive, especially for high-dimensional data. Hybrid strategies that combine symbolic regression with statistical regularization or neural network embedding's may offer practical scalability while preserving interpretability.

D. Uncertainty and Causality in Interpretable AI

Statistical interpretability is incomplete without uncertainty quantification. Bayesian generalized additive models allow posterior intervals for each function f_j(x_j), giving practitioners not only point estimates but also confidence bounds. In high-stakes applications like oncology prognosis, uncertainty quantification helps clinicians weigh the reliability of model predictions before making treatment decisions [48].

Causal interpretability extends these insights further. Structural causal models can disentangle spurious correlations from true causal effects. For instance, in socioeconomic data, income may correlate with health outcomes, but causal models can separate direct effects (e.g., access to healthcare) from confounding pathways [49]. Embedding causal structure into interpretable AI models ensures that transparency extends beyond correlation into mechanism-based explanation.

E. Mechanistic Interpretability of Neural Networks

Recent progress in mechanistic interpretability has shown that certain neurons in language models align with human-interpretable features such as sentiment or syntactic structure [50].

These findings suggest that deep networks may not be inherently opaque but instead require systematic analytical frameworks to uncover their internal logic. While still labor-intensive, this research points toward a long-term synthesis of black-box AI with the interpretability ethos of statistics.

F. Practical Implications and Challenges

The comparative review highlights several practical takeaways:

1. Accuracy vs. Interpretability – While deep learning often achieves the highest accuracy, interpretable models such as EBMs can approach comparable performance in many real-world datasets.
2. Trust and Accountability – In domains regulated by law or ethics (e.g., healthcare, credit scoring), inherently interpretable models may be preferable even at a modest cost in accuracy.
3. Computational Burden – Methods like symbolic regression and Shapley value attribution can be computationally intensive, limiting their scalability.
4. Approximation Risks – Post-hoc methods, though useful, must be interpreted cautiously as they may not reflect the true decision boundaries of the underlying model.

Together, these results underscore that bridging classical statistics and modern AI is less about replacing one paradigm with the other and more about integrating complementary strengths.

Table 4: Application Domains and Interpretability Needs

Domain	Key Requirement	Preferred Model Characteristics
Healthcare	Transparency, uncertainty quantification	Glass-box models (GAM, EBM), Bayesian models
Finance	Accountability, fairness	SHAP/LIME explanations, causal modeling
Policy	Simplicity, reproducibility	Linear/GLM with interpretable extensions
Engineering/Science	Mechanistic discovery	Symbolic regression, causal DAGs

V. CONCLUSION AND FUTURE WORK

This paper has examined the intersection of classical statistics and modern AI, emphasizing the importance of interpretability in data-science models. Classical approaches such as linear regression and generalized linear models offer clarity, uncertainty quantification, and rigorous inference. Modern AI, in contrast, excels in predictive accuracy and flexibility but often functions as a "black box."

Table 5: Comparison of Classical Statistics and Modern AI Models

Aspect	Classical Statistics (e.g., Regression, GLM)	Modern AI (e.g., DNN, Ensembles)
Interpretability	High — coefficients and functional forms directly interpretable	Low — parameters difficult to interpret
Uncertainty	Explicit (confidence intervals, hypothesis tests)	Often absent, unless Bayesian methods applied
Data Needs	Performs well on small to moderate datasets	Requires large datasets for stability
Flexibility	Limited — mainly linear or specified non-linear forms	Very flexible — captures complex, non-linear patterns
Transparency	Transparent assumptions, explainable outputs	Opaque, requires post-hoc explanations

The methods reviewed including generalized additive models, explainable boosting machines, SHAP values, LIME, symbolic regression, Bayesian modeling, and mechanistic interpretability illustrate that these paradigms are not mutually exclusive. Instead, they can be integrated to create models that balance accuracy with transparency.

The comparative discussion highlighted several key insights. First, interpretable glass-box models can achieve accuracy close to that of deep learning methods in many applied domains, reducing the need for full reliance on opaque architectures. Second, post-hoc explanation techniques provide useful approximations but must be interpreted cautiously, as they may not perfectly reflect the true internal logic of the model. Third, methods such as symbolic regression and mechanistic interpretability point toward a long-term synthesis, in which models are not only high-performing but also expressed in human-understandable mathematical form.

Several directions remain open for future research. Hybrid frameworks that embed statistical constraints into machine-learning models represent a promising path, particularly for healthcare, climate modeling, and finance where interpretability is indispensable. Advances in scalable symbolic regression and causal machine learning could further strengthen the bridge between statistics and AI. Finally, the growing emphasis on ethics, governance, and regulation of AI suggests that interpretable models will play an increasingly central role, not only as technical artifacts but as instruments of accountability.

Table 6: Open Challenges in Interpretable AI

Challenge	Description	Research Directions
Scalability	Interpretability methods often struggle on very large datasets	Approximation techniques, distributed implementations
Faithfulness	Post-hoc explanations may not reflect model internals	Inherently interpretable architectures
Evaluation Metrics	Lack of consensus on measuring interpretability	Human-centered evaluation frameworks
Adversarial Robustness	Explanations can be manipulated	Robust attribution and defense strategies
Causal Interpretability	Current models often descriptive, not causal	Integration with causal inference frameworks

The path forward lies not in replacing one paradigm with the other but in creating a cohesive framework that unites the inferential rigor of statistics with the representational power of AI. Achieving this balance will be essential for building trustworthy, transparent, and actionable data-science systems.

Figure 9: Conceptual Roadmap Flowchart

REFERENCES

[1] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," Digital Signal Processing, vol. 73, pp. 1–15, 2018.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[3] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[4] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," J. Royal Statistical Society A, vol. 135, no. 3, pp. 370–384, 1972.
[5] D. Freedman, "Statistical models and shoe leather," Sociological Methodology, vol. 21, pp. 291–313, 1991.
[6] H. Caruana et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," Proc. 21st ACM SIGKDD Int. Conf., pp. 1721–1730, 2015.
[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," Proc. 22nd ACM SIGKDD Int. Conf., pp. 1135–1144, 2016.
[8] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774, 2017.
[9] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," Science, vol. 324, no. 5923, pp. 81–85, 2009.
[10] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, and L. Schubert, "Zoom in: Circuits in deep neural networks," Distill, 2020.
[11] G. A. F. Seber and A. J. Lee, Linear Regression Analysis, 2nd ed., Wiley, 2012.
[12] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," J. Royal Statistical Society A, vol. 135, no. 3, pp. 370–384, 1972.
[13] T. Hastie and R. Tibshirani, Generalized Additive Models, CRC Press, 1990.
[14] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed., Springer, 2009.
[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
[16] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105, 2012.
[17] C. Molnar, Interpretable Machine Learning, 2nd ed., Leanpub, 2022.
[18] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, 2018.
[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," KDD, 2016.
[20] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," NeurIPS, 2017.
[21] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," AAAI Conf. Artificial Intelligence, pp. 1356–1363, 2020.
[22] H. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," J. Royal Statistical Society B, vol. 82, no. 4, pp. 1059–1086, 2020.
[23] A. Saltelli, K. Chan, and E. Scott, Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models, Wiley, 2004.
[24] H. Caruana et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital readmission," KDD, 2015.
[25] M. A. Hayat, S. A. Ahmed, S. Fatima, E. F. Irfan, M. O. Nizamaniand A. Khalil, "TINY MACHINE LEARNING (TINYML) ADVANCEMENTS FOR INTELLIGENT BATTERY-POWERED IOT SENSORS", TINY MACHINE LEARNING (TINYML) ADVANCEMENTS FOR INTELLIGENT BATTERY-POWERED IOT SENSORS, vol. 3, no. 8, pp. 818–832, Aug. 2025, doi: 10.5281/zenodo.16932733.
[26] C. Olah et al., "Zoom in: Circuits in deep neural networks," Distill, 2020.
[27] Z. Lipton, "The mythos of model interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.
[28] R. Doshi-Velez and F. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[29] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," International Data Privacy Law, vol. 7, no. 2, pp. 76–99, 2017.
[30] T. Hastie and R. Tibshirani, Generalized Additive Models, CRC Press, 1990.
[31] H. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital readmission," KDD, 2015.
[32] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games, vol. 2, H. W. Kuhn and A. W. Tucker, Eds. Princeton Univ. Press, 1953, pp. 307–317.
[33] M. Siddiqui, H. M. A. Siddiqui, S. Hussain, Dr. Muhammad Faseeh Ullah Khan, Syed Faraz Aliand M. A. Hayat, "INTERACTION OF FINANCIAL LITERACY IN IMPULSIVE BUYING BEHAVIOR THEORY", Remittances Review, vol. 9, no. 3, pp. 780–802, 2024, doi:
[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," KDD, 2016.
[35] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," AAAI, 2020.
[36] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," Science, 2009.
[37] J. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, 1992.
[38] A. Gelman, J. B. Carlin, H. S. Stern, D. Dunson, A. Vehtari, and D. B. Rubin, Bayesian Data Analysis, 3rd ed., Chapman and Hall/CRC, 2013.
[39] M. A. Hayat, S. Ahmed, M. R. K. Khan, E. M. Zaka, E. F. Irfanand E. R. Zaka, "Blockchain-Secured Iot Framework for Smart Waste Management in Urban Environments", The Critical Review of Social Sciences Studies, vol. 3, no. 3, Aug. 2025, doi: 10.5281/zenodo.17079639.
[40] C. Olah et al., "Zoom in: Circuits in deep neural networks," Distill, 2020.
[41] H. Caruana et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," KDD, 2015.
[42] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," Digital Signal Processing, 2018.
[43] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine learning with neural nets," Advances in Neural Information Processing Systems (NeurIPS), 2021.
[44] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," NeurIPS, 2017.
[45] D. Slack et al., "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," AAAI, 2020.
[46] E. F. Irfan, E. R. Zaka, E. S. Rehman, B. Sattar, S. A. Haiderand M. A. Hayat, "An IOT-Driven Smart Agriculture Framework for Precision Farming, ResourceOptimization, and Crop Health Monitoring", ACADEMIA International Journal for Social Sciences, vol. 4, no. 3, pp. 3329–3342, Aug. 2025, doi: 10.63056/ACAD.004.03.0615.
[47] A. Cranmer, P. Sanchez-Gonzalez, P. Battaglia, and D. Spergel, "Discovering symbolic models from deep learning with inductive biases," NeurIPS, 2020.
[48] A. Gelman et al., Bayesian Data Analysis, 3rd ed., Chapman & Hall/CRC, 2013.
[49] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge Univ. Press, 2009.
[50] C. Olah et al., "Zoom in: Circuits in deep neural networks," Distill, 2020.
[51] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," NeurIPS, 2017.
[52] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," Science, 2009.
[53] E. R. Zaka, S. M. Mushtaher Uddin, M. A. Hayat, A. Murtaza, S. A. Haiderand C. lal Beejal, "AI-Driven Cybersecurity for IoT–Cloud Ecosystems", Physical Education, Health and Social Sciences, vol. 3, no. 3, Sep. 2025, doi: 10.5281/zenodo.17079810.

[54] L. Saeed, R. Khan, D. S. Ali Durrani, C. Y. Mehmoodand M. A. Hayat, "HR Beyond the Office: Leveraging AI to Lead Distributed Teams and Cultivate Organizational Culture in the Age of Remote and Hybrid Work", ACADEMIA International Journal for Social Sciences (AIJSS), vol. 4, no. 3, pp. 291–310, Jul. 2025, doi: 10.63056/ACAD.004.03.0361.

[55] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," Science, 2009.

[56] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed., Cambridge University Press, 2009.

[57] M. A. Hayat, S. Ahmed, M. R. K. Khan, E. M. Zaka, E. F. Irfanand E. R. Zaka, "Blockchain-Secured Iot Framework for Smart Waste Management in Urban Environments", The Critical Review of Social Sciences Studies, vol. 3, no. 3, Aug. 2025, doi: 10.5281/zenodo.17079639.