MANTECH Publications

Journal of Converging Technologies in Intelligent Computing
Volume 1, Issue 3, September-December 2025

Explainable Artificial Intelligence and Hybrid Machine Learning Models for Transparent Intelligent Decision Support: A Structured Review of Methods, Architectures, and Deployment Challenges

Ananya R. Verma[1], Rohit K. Sahu[2], Priya M. Tiwari[3], Deepak S. Yadav[4]

ABSTRACT

Machine learning models now inform decisions in clinical triage, credit underwriting, and industrial maintenance, yet the most accurate models are often opaque, which limits accountability, regulatory compliance, and user trust. This review examines how Explainable Artificial Intelligence (XAI) and hybrid machine learning architectures can be combined to provide transparent intelligent decision support. A structured narrative review of representative literature from 2015 to 2025 was conducted and coded on four dimensions: explanation mechanism, model architecture, evaluation level, and application domain. The synthesis shows that post-hoc attribution methods such as LIME and SHAP dominate practice [1], [2], but their fidelity and stability are conditional and can be undermined [3], [4]. Hybrid designs that pair an interpretable glass-box model with a high-capacity learner, or that route predictions through human-understandable concepts, narrow the accuracy–transparency gap while keeping explanations faithful [5], [6]. The paper proposes a layered reference framework, identifies five recurring hybrid design patterns, maps them to clinical, financial, and industrial contexts, and outlines open problems in evaluation, causal grounding, and governance. The evidence supports hybrid-by-design transparency for high-stakes decision support.

KEYWORDS: *Explainable artificial intelligence, Hybrid machine learning, Decision support systems, Interpretability, SHAP, LIME, Glass-box models, Neurosymbolic AI, Trustworthy AI, Counterfactual explanations*

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

INTRODUCTION

Intelligent decision support systems increasingly rely on machine learning to rank patients for triage, approve or decline credit, and schedule maintenance on critical equipment. The models that achieve the highest accuracy on complex, high-dimensional data, namely deep neural networks and large tree ensembles, are also the least transparent. Their internal logic cannot be inspected directly by the clinician, loan officer, or plant engineer who must act on the output. This opacity is not cosmetic. A widely cited clinical study showed that a high-performing pneumonia-risk model had learned that patients with asthma were at lower risk, because such patients historically received more intensive care; the pattern was statistically valid but would have been dangerous if used to withhold treatment [5]. Errors of this kind are difficult to detect when only accuracy metrics are examined.

Two schools of thought have emerged. The first accepts black-box accuracy and adds post-hoc explanation, approximating or probing the trained model after the fact [7], [8]. The second argues that for high-stakes decisions, inherently interpretable models should replace black boxes altogether, because an explanation that is not the model itself can never be fully faithful [9]. Hybrid machine learning, which combines interpretable and high-capacity components within one pipeline, offers a pragmatic middle path, yet its relationship with explanation techniques has rarely been examined systematically. Regulation adds urgency: the European Union Artificial Intelligence Act imposes transparency and human-oversight obligations on high-risk systems, including those used in credit scoring and medical contexts [10].

The problem addressed here can be stated precisely. Practitioners building decision support systems lack consolidated, evidence-based guidance on which combination of explanation technique and hybrid architecture delivers faithful, stable, and usable transparency for a given risk level and user group. This paper synthesises the literature to answer that question and proposes a layered reference framework, following the sections listed in the journal structure: related work, research gap, objectives, methodology, findings, discussion, limitations, future scope, and conclusion.

LITERATURE REVIEW

Research on explainability has grown rapidly since 2015 and can be organised into foundations, post-hoc methods, and intrinsically interpretable or hybrid models.

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Foundations Of Explainability

Doshi-Velez and Kim distinguished interpretability from mere performance and proposed three evaluation levels: functionally-grounded, human-grounded, and application-grounded [11]. Arrieta et al. extended this into a taxonomy linking explainability to responsible AI and to the needs of different audiences [7]. Guidotti et al. organised the field around four problems: explaining a model, explaining an individual outcome, inspecting a model, and designing a transparent model directly [8]. Together these works establish that explainability is audience- and task-dependent, a premise that underlies the framework proposed here.

Post-Hoc Explanation Methods

LIME explains an individual prediction by fitting a sparse linear surrogate to perturbed samples around it [1]. SHAP unified several attribution methods under Shapley-value theory and provided additivity and consistency guarantees [2]. TreeSHAP later made exact Shapley attribution tractable for tree ensembles and showed how many local explanations can be aggregated into global insight [12]. These methods are model-agnostic or nearly so and easy to deploy, which explains their dominance. Their reliability, however, has been questioned. Sanity checks revealed that some saliency methods produce nearly identical maps even when model parameters are randomised [3], and adversarial classifiers can be constructed that behave fairly on the perturbed samples used by LIME and SHAP while remaining biased on real data [4]. Post-hoc output must therefore be audited, not trusted by default.

Interpretable And Hybrid Models

Intelligible generalised additive models with pairwise interactions achieved accuracy comparable to complex learners on healthcare tasks while exposing every learned term for expert review, which allowed the asthma artefact to be identified and corrected [5]. The Explainable Boosting Machine, distributed in the InterpretML toolkit, made this family practical for tabular data [13]. Concept bottleneck models route predictions through human-defined concepts, so clinicians can inspect and even intervene on intermediate concepts at test time [6]. Neurosymbolic approaches integrate neural learning with symbolic knowledge, promising explanations expressed in domain rules [14]. In finance, Shapley-based explanations have been combined with network analysis to support credit-risk scoring on lending-platform data [15]. Collectively these studies show that transparency can be engineered into the architecture rather than appended to it.

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

RESEARCH GAP

Existing surveys catalogue explanation techniques extensively [7], [8], but three gaps persist. First, hybrid architectures are usually treated as secondary to the explanation method, so their transparency properties are not compared systematically. Second, evaluation is fragmented: fidelity, stability, and user comprehension are rarely assessed together, and the manipulability of explanations is seldom considered in design [3], [4]. Third, little work maps explanation properties to the requirements of specific decision contexts, such as clinical, financial, and industrial settings, and to emerging regulation [10]. This review addresses these gaps jointly.

OBJECTIVES

- To classify explanation techniques and hybrid architectures relevant to intelligent decision support.
- To compare them on fidelity, stability, computational cost, and decision-maker usability.
- To propose a layered reference framework combining hybrid models with an explanation and audit layer.
- To map design patterns to application domains and identify open research problems.

METHODOLOGY

Review Design And Source Selection

A structured narrative review was adopted, which suits a field where heterogeneous methods and metrics preclude quantitative pooling. Literature was drawn from IEEE Xplore, the ACM Digital Library, ScienceDirect, SpringerLink, and arXiv, covering publications from 2015 to 2025 together with foundational earlier work. Search strings combined explainable AI, interpretable machine learning, hybrid model, neurosymbolic, and decision support with the domain terms clinical, credit, and maintenance. Studies were retained if they introduced or evaluated an explanation technique or hybrid architecture with a reproducible description and relevance to decision support. Purely theoretical papers without an implementable method were excluded, except landmark position papers that shaped the field [9].

Analytical Framework

Each retained work was coded on four dimensions: explanation mechanism, model architecture, evaluation level following Doshi-Velez and Kim [11], and application domain. Figure 1 shows the taxonomy used for coding. Comparative ratings in Tables 1 to 3 are qualitative judgements (High, Medium, Low) made by the authors from properties and empirical observations reported in the cited studies; no new experiments were performed. Table

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

1 compares the principal explanation techniques on scope, model access, fidelity, stability, and computational cost.

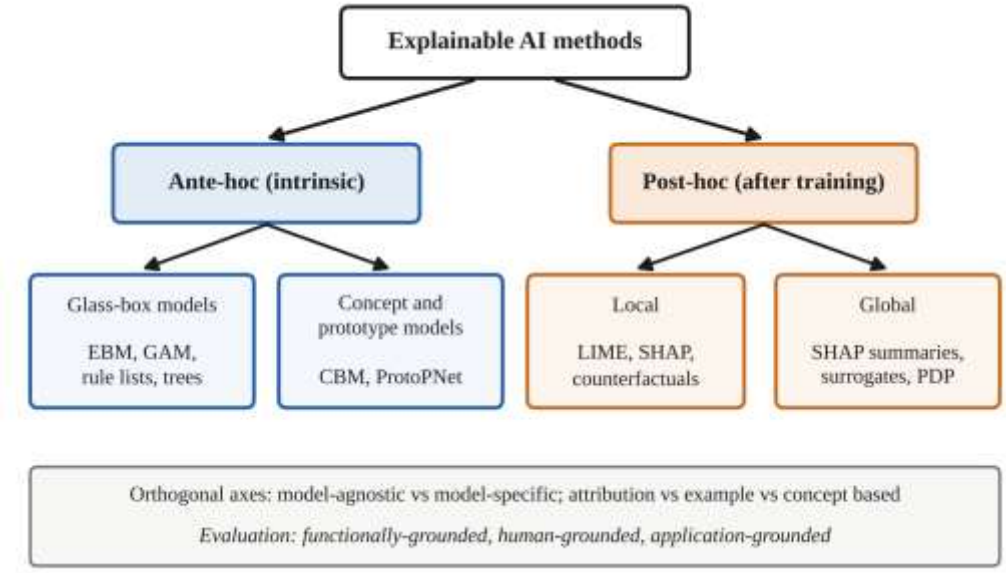


Figure 1. Taxonomy of explainable AI methods used for coding: ante-hoc versus post-hoc approaches, with local and global scopes and the three evaluation levels

Table 1. Comparative Properties Of Principal Explanation Techniques (Qualitative Ratings)

Technique	Scope	Model Access	Fidelity	Stability	Cost
LIME [1]	Local	Agnostic	Medium	Low–Medium	Medium
Kernel SHAP [2]	Local, global	Agnostic	High	Medium	High
TreeSHAP [12]	Local, global	Tree models	High	High	Low
Saliency maps [3]	Local	Neural networks	Variable	Low–Medium	Low
Counterfactuals [7]	Local	Agnostic	Medium	Medium	Medium
EBM / GAM [5], [13]	Local, global	Intrinsic	Exact	High	Low

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

RESULTS AND FINDINGS

Finding 1: post-hoc attribution dominates, but fidelity is conditional. In the reviewed literature, SHAP- and LIME-style attribution appears as the most common route to explanation because it attaches to any trained model [1], [2]. Table 1 shows why the choice matters: TreeSHAP combines high fidelity with low cost for tree ensembles, whereas kernel-based SHAP and LIME become expensive or unstable as feature dimensionality grows, and saliency maps require sanity testing before use [12], [3]. Glass-box models avoid the fidelity question entirely because the explanation is the model [5], [13].

Finding 2: a layered hybrid framework. Synthesising the reviewed designs yields the reference framework in Figure 2. Data enter a predictive stage where an interpretable glass-box model and a high-capacity model operate in parallel. A hybrid arbitration layer combines them through residual learning or confidence-gated routing: the glass-box model handles cases in which it is confident, and the high-capacity model handles the remainder or corrects the glass-box residual. The arbitrated output passes to a transparency stage that generates attribution, counterfactual, and concept-level explanations before the human decision-maker sees it. A feedback path logs explanations, monitors fidelity, drift, and subgroup fairness, and returns findings to the data layer. The key design property is that every decision carries an explanation whose faithfulness is known: exact for the glass-box component and audited for the residual component.

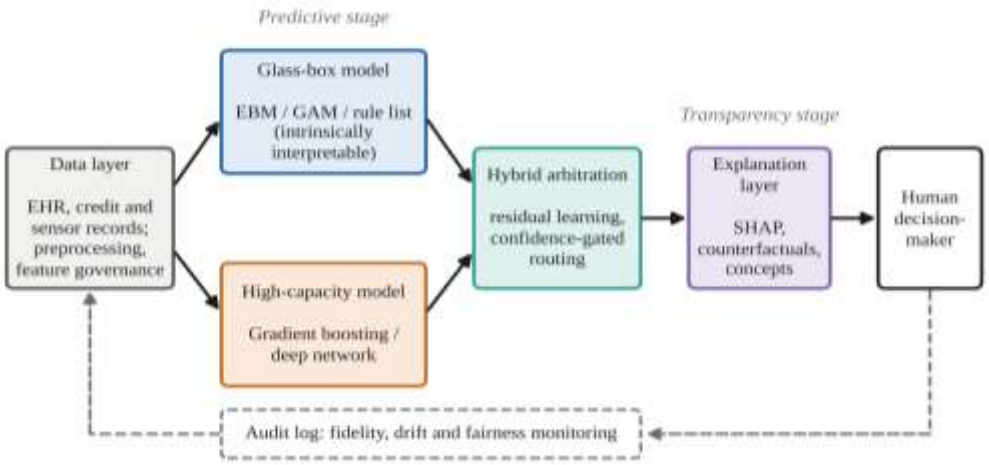


Figure 2. Proposed layered reference framework for transparent intelligent decision support combining a glass-box model, a high-capacity model, hybrid arbitration, an explanation layer, and an audit feedback loop

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Finding 3: five recurring hybrid patterns. Table 2 summarises the patterns identified. Their common aim is to confine opacity to the smallest part of the pipeline that accuracy requires. Residual and gated designs preserve exact explanations for most predictions, concept bottlenecks make intermediate reasoning clinician-readable, and neurosymbolic designs express explanations as domain rules [5], [13], [6], [14]. Wrapping a tree ensemble with TreeSHAP remains the lowest-effort option but leaves the explanation separate from the model [12].

Table 2. Hybrid Architecture Patterns For Transparent Decision Support

Pattern	Mechanism	Transparency Gain	Main Limitation	Evidence
Glass-box plus residual	Additive model first; boosted learner fits residual	Main effects exactly readable	Residual needs post-hoc attribution	[5], [13]
Confidence-gated routing	Simple model decides confident cases; complex model takes the rest	Most decisions fully transparent	Gate calibration is critical	Author synthesis
Concept bottleneck	Input, then human concepts, then label	Clinician-readable; test-time intervention	Needs concept annotations	[6]
Neurosymbolic	Neural learning coupled with symbolic rules	Rule-form explanations	Knowledge-engineering cost	[14]
Post-hoc wrapped ensemble	Boosted trees plus TreeSHAP	Low-effort attribution	Explanation separate from model	[12]

Finding 4: the accuracy–transparency trade-off is weaker for hybrids. Figure 3 positions model families on qualitative scales derived from reported behaviour. Classical models lie along a frontier on which gaining predictive capacity costs interpretability. Reports on additive models with interactions, explainable boosting, and concept bottlenecks indicate predictive performance competitive with standard learners on the tasks studied [5], [13], [6], which places

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

them above that frontier. The placement is illustrative rather than measured, and the gap depends on dataset complexity; on perceptual data with very high-dimensional raw inputs, glass-box models remain less competitive.

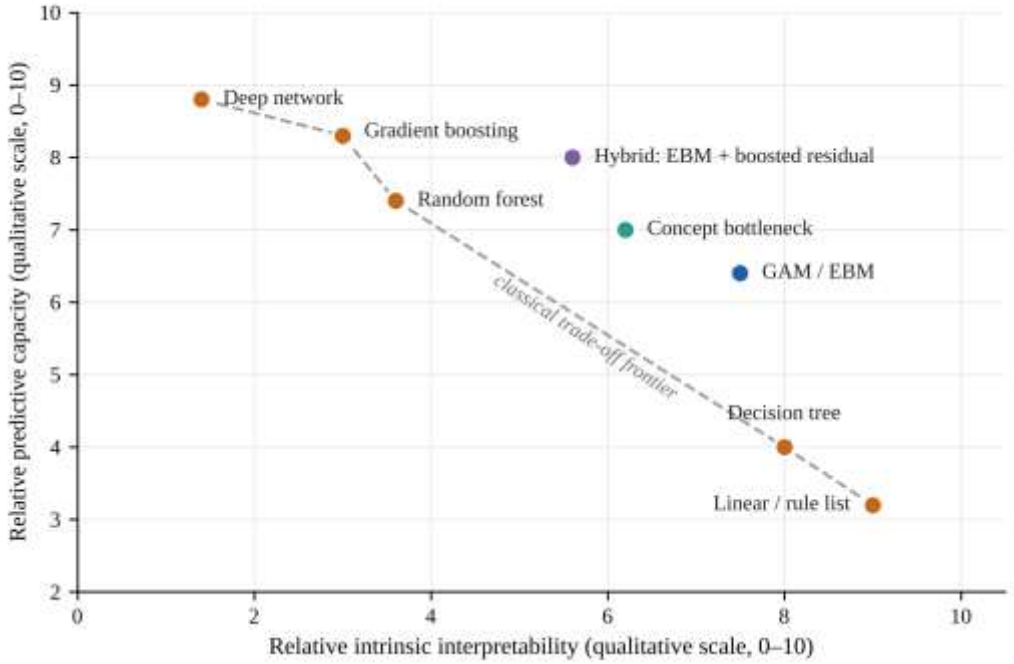


Figure 3. Qualitative positioning of model families on interpretability and predictive capacity; hybrid and concept-based designs sit above the classical frontier (illustrative, not measured)

Finding 5: explanations must themselves be validated. Manipulable or insensitive explanations [3], [4] make an audit layer a core component rather than an optional extra. In our reading, functionally-grounded proxies such as fidelity scores were reported more often than application-grounded studies with domain experts, even though the latter are the stronger evidence of decision-support value [11]. Table 3 maps decision contexts to recommended patterns.

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Table 3. Decision Contexts And Recommended Explanation Design Patterns

Domain	Principal Explanation Need	Recommended Pattern	Governance Driver
Clinical decision support	Expert-editable risk terms; detection of spurious effects	Glass-box plus residual; concept bottleneck [5], [6]	Patient safety; human oversight [10]
Credit risk and lending	Per-applicant reasons; fairness audit	Glass-box plus residual with Shapley reasons [15]	Transparency obligations [10]
Predictive maintenance	Actionable what-to-change guidance	Gated routing with counterfactuals [7]	Operational safety and reliability
Public-service eligibility	Contestable, rule-form reasons	Neurosymbolic or glass-box [14], [9]	Accountability and appeal rights

DISCUSSION

Technical Interpretation

Hybrid designs work because predictive structure is unevenly distributed. In many tabular domains most signal resides in low-order effects, which additive glass-box terms capture exactly, while a smaller portion arises from higher-order interactions that justify a high-capacity learner. Decomposing the model accordingly keeps the explanation of the dominant part exact and confines approximate attribution to the residual. This supports the position that interpretable models should be the default for high-stakes use [9], while acknowledging the pragmatic case for black boxes where perceptual complexity demands them [7]. The distinction between plausible and faithful explanations is central: an explanation that merely appears reasonable to the user can still misdescribe the model, which is why fidelity testing and manipulation audits are needed [3], [4].

Engineering And Scientific Significance

From an engineering perspective, the framework treats transparency as a design requirement with testable properties rather than a reporting step after deployment. Modular components can be validated independently, and logged explanations create an audit trail that supports the transparency and human-oversight obligations of the EU Artificial Intelligence Act [10]. Scientifically, the review shows that progress depends on measurable explanation quality:

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

standard fidelity, stability, and comprehension metrics, evaluated at all three levels proposed by Doshi-Velez and Kim [11], would convert XAI from an art into an empirical discipline.

Practical Applications

In clinical decision support, glass-box components let experts inspect and edit learned risk terms, preventing artefacts such as the asthma effect [5]. In credit risk, adverse-action reasons and Shapley-based drivers can be generated for each applicant while a boosted residual preserves discrimination power [15]. In predictive maintenance, counterfactual explanations tell operators which sensor conditions would have to change to avoid a failure alert. Practitioners should begin with a glass-box baseline, adopt a high-capacity component only when the measured gain justifies the added audit burden, and log every explanation served to a decision-maker.

LIMITATIONS

- The review is narrative and structured rather than a full systematic meta-analysis; heterogeneous datasets and metrics prevent quantitative pooling of results.
- The ratings in Tables 1 to 3 and the positioning in Figure 3 are qualitative syntheses by the authors and should be treated as hypotheses to test, not measured benchmarks.
- Selection from five English-language sources may introduce publication and language bias, and explanations for language-model-based decision support are only lightly covered.
- The proposed framework is conceptual and has not been validated empirically within this paper.

FUTURE SCOPE

- Benchmark the five hybrid patterns on shared tabular datasets with standardised fidelity, stability, and comprehension metrics.
- Integrate causal models and counterfactual reasoning so that explanations recommend feasible actions rather than only describe correlations.
- Conduct application-grounded user studies with clinicians, credit officers, and maintenance engineers to measure effects on decision quality and appropriate trust.
- Develop lightweight explanation and audit tooling aligned with regulatory documentation requirements, including edge deployment.

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

CONCLUSION

Opaque models cannot be accepted uncritically in decision support where errors affect health, finance, and safety. This review found that post-hoc attribution methods are valuable but conditional on fidelity and stability, whereas hybrid architectures that keep most of the decision logic in interpretable components, supplemented by an audited explanation layer, offer the strongest balance of accuracy and transparency. The proposed layered framework, five hybrid design patterns, and domain mapping give practitioners actionable guidance. Realising their promise requires standardised evaluation, causal grounding, and human-centred validation, which define the agenda for future work.

REFERENCES

1. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2016; 1135–1144.
2. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30.* 2017; 4765–4774.
3. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems 31.* 2018; 9505–9515.
4. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 2020; 180–186.
5. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2015; 1721–1730.
6. Koh PW, Nguyen T, Tang YS, et al. Concept bottleneck models. *Proceedings of the 37th International Conference on Machine Learning (PMLR 119).* 2020; 5338–5348.
7. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion.* 2020; 58: 82–115.

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

8. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Computing Surveys.* 2018; 51(5): Article 93.
9. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence.* 2019; 1(5): 206–215.
10. European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union.* 2024.
11. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.* 2017.
12. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence.* 2020; 2(1): 56–67.
13. Nori H, Jenkins S, Koch P, Caruana R. InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223.* 2019.
14. Garcez A d'Avila, Lamb LC. Neurosymbolic AI: The 3rd wave. *arXiv preprint arXiv:2012.05876.* 2020.
15. Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable machine learning in credit risk management. *Computational Economics.* 2021; 57: 203–216.

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved

***Author for Correspondence:**
Ananya R. Verma
E-mail: ananya.verma.cse@gmail.com

[1]Associate Professor, Department of Computer Science and Engineering, Government Engineering College, Bilaspur, Chhattisgarh
[2]Assistant Professor, Department of Computer Science and Engineering, Government Engineering College, Bilaspur, Chhattisgarh
[3]Assistant Professor, Department of Computer Science and Engineering, Government Engineering College, Bilaspur, Chhattisgarh
[4]M.Tech Scholar, Department of Computer Science and Engineering, Government Engineering College, Bilaspur, Chhattisgarh

Received Date: December 03, 2025
Accepted Date: December 16, 2025
Published Date: December 27, 2025

Citation: Ananya R. Verma, Rohit K. Sahu, Priya M. Tiwari, Deepak S. Yadav. Explainable Artificial Intelligence and Hybrid Machine Learning Models for Transparent Intelligent Decision Support: A Structured Review of Methods, Architectures, and Deployment Challenges. Journal of Converging Technologies in Intelligent Computing. 2025; 1(3): 204-216p.

Page 204-216 © MANTECH PUBLICATIONS 2025. All Rights Reserved