Review

An Overview of Recent Interpretability and Explainability Approaches for Tree-Based Ensembles

Alexandros Miteloudis * and Ioannis Hatzilygeroudis *

Posted Date: 30 April 2026

doi: 10.20944/preprints202604.2101.v1

Keywords: random forest; gradient boosting machine; decision tree based ensembles; interpretability models; explainability models

 doi:10.20944/preprints202604.2101.v1

Review

An Overview of Recent Interpretability and Explainability Approaches for Tree-Based Ensembles

Alexandros Miteloudis [1,*] and Ioannis Hatzilygeroudis [1,2,*]

[1] School of Science and Technology, Hellenic Open University, 26335 Patras, Greece;

[2] Department of Computer Engineering & informatics, University of Patras, 26500 Patras, Greece;

* Correspondence: almiteloudis@gmail.com (A.M.); ihatz@ceid.upatras.gr (I.H.)

Abstract

Decision tree ensembles, such as Random Forests and Gradient Boosting Machines, achieve high predictive accuracy but often suffer from limited transparency due to their structural complexity. Due to this lack, interpretability challenges arise in domains where model understanding, accountability, and trust are essential. So, many interpretability/explainability techniques have been proposed for tree-based ensembles. However, although there are enough surveys or overviews concerning interpretability/explainability in artificial intelligence or machine learning in general, there are very few surveys of overviews on interpretability/explainability for tree-based ensembles. This paper provides an overview of recent approaches to interpretability and explainability in decision tree ensembles. We present two categorizations; one based on the kind of technique/architecture used and the second based on the level of scope. The former is a unified taxonomy of acquired (or post-hoc) and inherent methods further analyzed in two more levels. The latter concerns the distinction between local (or instance-related) and global (or model-related) methods. We additionally provide a survey of the interpretability/explainability methods/techniques used in various domain applications, like healthcare, finance, law, privacy preserving. This overview clarifies the current landscape of interpretable/explainable ensemble learning, explicitly addressing emerging challenges. Ultimately, it aims to support researchers and practitioners in selecting and developing ensemble models that move beyond the traditional accuracy-interpretability trade-off, aligning predictive power with strict regulatory, operational, and domain-specific transparency requirements.

Keywords: random forest; gradient boosting machine; decision tree based ensembles; interpretability models; explainability models

1. Introduction

Decision trees (DTs) have been a cornerstone of machine learning since the pioneer works on Classification and Regression Trees (CART) by Breiman et al. [1] and Quinlan's C4.5 algorithm [2]. Their hierarchical, rule-based structure offers a uniquely transparent view of how input features map to predictions, enabling model validation by domain experts in various application domains, like medicine, finance, and law. This inherent interpretability of DTs has led to easy production of explanations of the decisions made based on them. However, the predictive performance of single trees weakens quickly on large or noisy data resulting in reduced accuracy. Therefore, tree-based ensembles (TBEs), like Random Forests [3], Gradient Boosting Machines [4] and their modern variants, have been devised to regain accuracy by combining a multitude of (weak) trees. However, the complexity of the resulted model buries the decision logic inside forests of branching paths, thus reducing or even vanishing interpretability and hence explainability. This is a problem not only for DTs but in general for AI approaches or techniques that act like a black box [5,6]. This loss of transparency hinders not only human understanding but also model debugging, accountability, and trust, particularly when automated decisions affect individuals or critical infrastructure.

 doi:10.20944/preprints202604.2101.v1

Given the above, the past decade has witnessed a new wave of research in making TBEs, and complex artificial intelligence (AI) models in general, interpretable and explainable. General surveys or reviews on explainable AI (XAI) [7–9] provide broad taxonomies of interpretable and/or explainable models and techniques, generally for machine learning approaches, including references to TBEs too. However, by being so broad they lack depth on dealing with TBEs. Only few surveys that focuse on interpretability and explainability for TBEs have been emerged, such as Haddouchi and Berrado [3], Gonçalves and Carvalho [4], and Sepioło and Ligęza [10]. The first two refer to specific TBE approaches, the Random Forest (RF) and the Gradient Boosting (GB) respectively. Only Sepioło and Ligęza [10] refer to tree-based ensemble models in general, but it does not include an adequately wide analysis and taxonomy of interpretability and/or explainability approaches or techniques for TBEs.

As a result, practitioners often face difficulties in selecting appropriate interpretability/explainability techniques when dealing with TBE architectures and domain-specific constraints. To fill in the above gap, we present in this paper an overview of recent (from 2020 to 2025) approaches or models concerning interpretability and explainability of TBEs. Compared with previous surveys, our work offers:

- A unified, cross-paradigm categorization of acquired and inherent interpretability or explainability methods. Unlike prior works that isolate techniques by specific ensemble types, this taxonomy goes beyond descriptive cataloging to map the underlying structural trade-offs applicable to all tree-based architectures.
- A survey of the specific interpretability and explainability approaches used in the above categories.
- A second taxonomy of interpretability and explainability methods based on their scope (global, local).
- A presentation of the interpretability and explainability methods/techniques used in various domain applications, like healthcare, finance, law, privacy preserving etc.
- Comparative analysis of the primary methodological models against four critical real-world constraints: Scalability, Computational Cost, Robustness/Stability, and Usability.
- Practical considerations for making possibly optimal design decisions for interpretable TBEs.
- A sketch addressing four critical open research challenges coming out of our analyses.

2. Background Knowledge

DTs are the base learners which TBEs are built on. DTs is one of the most significant models in machine learning due to their hierarchical, rule-based representation and way of working. The root and each internal node of a tree defines a decision condition on one feature, while each leaf represents a prediction outcome. The interpretability/explainability of a single tree arises from its transparent structure: a user can trace a complete decision path from root to leaf and understand precisely how each feature contributes to the final prediction [11].

To understand the complexity of modern ensembles, it is necessary to first define the foundations of the base learners and the aggregation mechanisms. In a typical classification task, we operate on a finite learning dataset, denoted as S, consisting of N samples. Each sample assigns a feature vector x_i to a specific class label $y_i \in M$, where M represents the label (class) set. Formally, the dataset can be represented as:

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \quad (1)$$

where x_i represents a d-dimensional feature vector:

$$x_i = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)}]^T \in X \quad (2)$$

where X represents the input feature space, involving d features:

$$F = \{F_1, F_2, \ldots, F_d\} \quad (3)$$

The fundamental objective of supervised learning is to induce a mapping function Ψ that accurately maps elements from the feature space X to the elements in the label set M:

 doi:10.20944/preprints202604.2101.v1

$$\Psi : X \rightarrow M \quad (4)$$

This function is constructed to satisfy desired performance criteria, such as maximizing accuracy or recall.

In the context of ensemble learning, rather than relying on a single function, the system aggregates the outputs of a set of k base classifiers to make a final decision:

$$\Pi = \{\Psi_1, \Psi_2, \ldots, \Psi_k\} \quad (5)$$

The final decision of the ensemble is then derived through aggregating, in some way, the outputs of those base models.

2.1. Tree Structure and Interpretability

At their core, decision trees operate by recursively partitioning the input feature space X into mutually exclusive rectilinear regions. The induction process typically follows a greedy, top-down approach, exemplified by algorithms such as CART [1] or C4.5 [2]. Starting from a root node representing the entire dataset, the algorithm searches for the optimal split, defined by a feature value threshold ($F_m \leq th$), that maximizes the homogeneity of the resulting child nodes. This homogeneity is quantified via impurity metrics, such as Gini impurity for classification or variance reduction for regression. The splitting process continues until a stopping criterion is met, such as reaching a maximum depth or a minimum sample size per leaf. Beyond their intuitive structure, decision trees exhibit several algorithmic properties that are directly tied to both their interpretability and their limitations. The greedy nature of the induction process implies that each split is optimized locally rather than globally. As a result, early splitting decisions strongly influence the final structure of the tree, often leading to suboptimal partitions when complex feature interactions exist. While this greedy induction strategy ensures computational efficiency, it also explains why small perturbations in the training data can lead to structurally different trees, contributing to model instability.

Another important characteristic of decision trees is their sensitivity to small perturbations in the training data. Minor changes in the dataset can result in important different split selections near the root, lead to structurally different trees with similar predictive performance. While this instability deals with generalization challenges, it also plays a key role in ensemble construction, where diversity among base learners is intentionally exploited to improve accuracy. From an interpretability perspective however, this variability highlights the fragility of explanations derived from individual trees.

Furthermore, the axis-aligned nature of standard decision tree splits imposes geometric constraints on the induced decision boundaries. Each split partitions the feature space along a single dimension, producing hyper-rectangular regions. Although this representation simplifies logical interpretation, it may require many nodes to approximate oblique or highly nonlinear boundaries. Consequently, model complexity often grows rapidly as predictive demands increase, creating tension between expressive power and human comprehensibility.

The inherent interpretability of a DT comes directly from its inference mechanism. To predict the outcome for an unseen instance, the model routes the data point from the root through a hierarchy of internal nodes. At each node, a specific Boolean test against a feature value attribute determines the subsequent path. The final prediction is yielded by the terminal leaf node reached, which aggregates the target values of the training samples falling into that region (e.g., via majority voting or averaging). Consequently, any prediction made by a DT can be explicitly traced as a unique conjunction of logical predicates, providing a transparent audit trail of the decision process.

Figure 1 illustrates this mapping between the hierarchical tree structure and the resulting geometric partition of a two-dimensional feature space. An input instance is routed from the root through a series of Boolean tests until a terminal leaf node is reached, providing a clear audit trail of the decision logic. Also, it is easy to produce a set of rules to model the decision tree:

 doi:10.20944/preprints202604.2101.v1

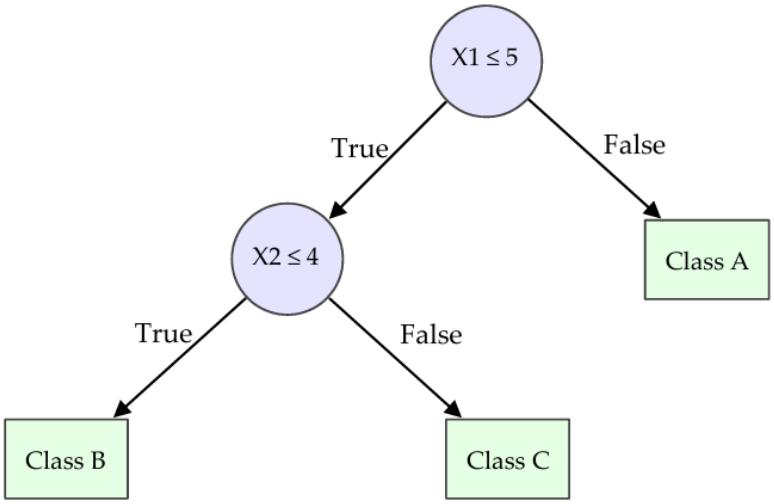

Figure 1. A visual representation of a decision tree and its inference flow.

If X1 > 5 then A, If X1 ≤ 5 and X2 ≤ 4 then B, If X1 ≤ 5 and X2 > 4 then C.

2.2. From Single Trees to Ensembles of Trees

Although single trees (such as CART [1] or C4.5 [2]) are intuitive, they often generalize poorly on complex or noisy data due to the variance described above. To overcome this limitation, ensemble methods combine many (weak) decision trees to achieve superior predictive accuracy.

Formally, given the ensemble of k decision trees defined previously as: $\Pi = \{\Psi 1, \Psi 2, \ldots, \Psi k\}$, each individual tree Ψj (for $j = 1, \ldots, k$) acts as a distinct mapping function $\Psi_j : X \rightarrow M$. When an unseen feature vector $x \in X$ is presented to the model, each tree independently computes a prediction $\Psi_j(x)$. The final output of the ensemble, denoted as $\Psi_{ens}(x)$, is determined by aggregating these individual predictions.

In *classification* tasks, this aggregation is typically achieved through majority voting, which can be mathematically expressed as:

$$\Psi_{ens}(x) = argmax_{c \in M} \sum_{j=1}^{k} \mathbb{I}\big(\Psi_j(x) = c\big) \quad (6)$$

where $\mathbb{I}(\cdot)$ is the indicator function that yields 1 if the condition is true (i.e., the *jth* tree predicts class c), and 0 otherwise.

In *regression* tasks, the aggregation is typically achieved by averaging the outputs of the individual trees, which mathematically expressed as:

$$\Psi_{ens}(x) = \frac{1}{k} \sum_{j=1}^{k} \Psi_j(x) \quad (7)$$

By aggregating multiple models, the ensemble effectively reduces the variance associated with individual trees, leading to a more robust and accurate decision boundary across the feature space X.

To thoroughly understand the above formulation, we define its components in the context of our established notation:

- $\Psi ens(x)$ represents the final ensemble output for input x.
- k is the total number of independent decision trees in the ensemble.
- $\Psi j(x)$ denotes the final prediction of the *jth* decision tree for input x.
- The summing term in formula in (6) computes the sums of the predictions of the individual trees for each class label and then the label with the largest sum is taken as the final output. What internally happens is that first each tree predicts a probability for each class label, then the average probability across all trees is computed and based on that the final class label (the one with the largest average value) is assigned to $\Psi j(x)$.

 doi:10.20944/preprints202604.2101.v1

– The formula in (7) computes the arithmetic mean of the continuous outputs of the k individual trees.

Two dominant paradigms define this landscape:

- *Bagging (Bootstrap Aggregating)*: Exemplified by Random Forests [3], this method builds multiple trees in parallel on different subsets of data and averages their predictions. Mathematically, for an ensemble of k trees (consistent with our previously defined ensemble Π), the final aggregated prediction $\Psi_{bag}(x)$ for an unseen feature vector $x \in X$ is defined by formulas (6) and (7).

Figure 2 shows how the training of a typical bagging model is performed and how its output is determined. *Sj* represents a *bootstrap sample*. Instead of training on the original dataset *S*, each tree *j* is trained on *Sj*, which is a dataset of the same size as the original, created by randomly sampling the original dataset *S with replacement*. So, all trees are trained in a parallel mode. The individual predictions of trees are aggregated to produce the final prediction $\Psi_{bag}(x)$. This aggregation reduces variance but creates a complex "forest" where individual paths are hard to untangle. Although individual trees in a Random Forest remain interpretable in isolation, the collective behavior of hundreds of such trees renders the overall decision process opaque.

- *Boosting*: Exemplified by Gradient Boosting Machines (GBMs) and variants like XGBoost [4], this method builds trees sequentially. Each new tree corrects the errors of the previous ones. In formal terms, this is an additive expansion where each new tree fits the negative gradient of the loss function. Mathematically, for an ensemble of k trees (consistent with our previously defined ensemble Π), the final aggregated prediction $\Psi_{boost}(x)$ for an unseen feature vector $x \in X$ is defined by formulas (5) and (6).

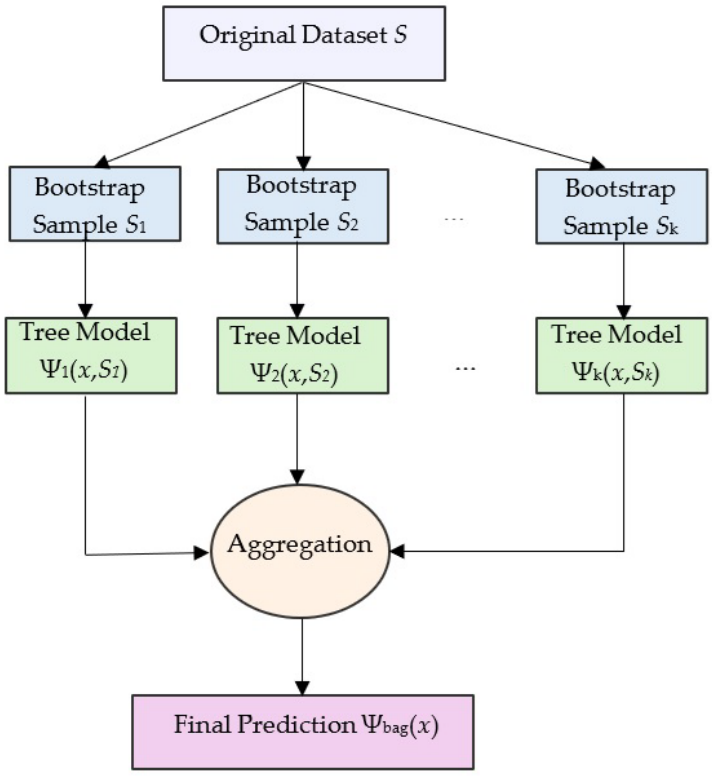

Figure 2. A visual representation of the Bagging (Bootstrap Aggregating) model architecture.

Figure 3 shows how the training of a typical boosting model is performed and how its output is determined. As it is clear, member trees are trained sequentially. Each *Sj* represents a *weighted sample*. Instead of training on the original dataset *S*, tree *j* is trained on *Sj*, which is a dataset of the same size as the original, created by assigning weights to instances. Misclassified instances in the previous stage are assigned larger weights, i.e., are given greater priority for correct prediction. The individual

 doi:10.20944/preprints202604.2101.v1

predictions of trees are aggregated to produce the final prediction $\Psi_{boost}(x)$. While highly accurate, the final prediction is a weighted sum of hundreds of additive terms. From an interpretability perspective, this sequential dependency makes it difficult to attribute predictions to a small, human-comprehensible set of rules.

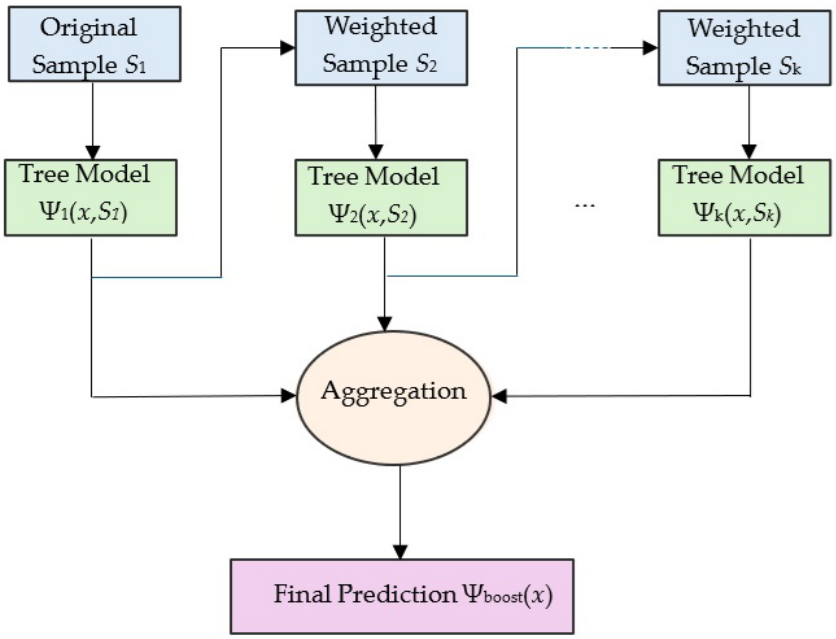

Figure 3. A visual representation of a typical Boosting model architecture.

While ensemble methods substantially improve predictive performance, they fundamentally alter the nature of the decision process. In contrast to a single tree, where a prediction corresponds to a single, traceable path, ensemble predictions emerge from the aggregation of many heterogeneous decision paths. This aggregation obscures the contribution of individual features and rules, making it difficult to associate a specific prediction with a concise logical explanation.

Importantly, the loss of interpretability is not merely a by-product of increased model size, but a consequence of how ensemble diversity is constructed. Techniques such as bootstrap sampling, random feature selection, and sequential residual fitting intentionally decorrelate individual trees. Although this decorrelation improves generalization, it also fragments decision logic across the ensemble, complicating any attempt to derive a unified explanation.

2.3. Interpretability vs Explainability

In the context of this paper, it is crucial to distinguish between two frequently confused terms: interpretability and explainability. Following recent XAI taxonomies [5,6,12], we define these as follows:

- *Interpretability*: Refers to models that are naturally understandable by humans without the need for secondary tools. The model's structure itself communicates the logic (e.g., a thin decision tree or a small rule list).
- *Explainability*: Refers to techniques applied to opaque models (black boxes) to approximate or visualize their decision-making process. For example, extracting feature importance scores from a dense Random Forest is an act of explainability, not interpretability.

In practice, these two notions often coexist within the same system, where inherently interpretable components are complemented by post-hoc explanatory tools.

While related, these concepts represent fundamentally different approaches to the "understanding criterion." Interpretability is a passive property of the model design, prioritizing

 doi:10.20944/preprints202604.2101.v1

structural simplicity (often at the potential cost of accuracy, though this trade-off is debated [13]). In contrast, Explainability is an active reconstruction process, prioritizing the preservation of the ensemble's high-dimensional performance while offering a more holistic view of its logic.

This distinction becomes particularly critical in the context of decision tree ensembles. While a single tree can often be interpreted directly by inspecting its structure, ensembles inherently violate the transparency criterion due to their distributed decision logic. As a result, most ensemble models cannot satisfy interpretability in the strict, intrinsic sense and instead rely on explainability mechanisms to approximate their behavior. Consequently, interpretability for ensembles is not an absolute property but a spectrum, ranging from fully opaque models with post-hoc explanations to constrained architectures designed to expose partial structural transparency.

Based on the above, the methodologies reviewed in this paper are largely separated by this distinction: whether they seek to extract logic from an opaque ensemble (acquired or post-hoc) or design an ensemble that is transparent by nature (inherent or ante-hoc).

Lim et al. [14] provided comparative studies demonstrating that while ensembles consistently outperform single trees in accuracy, the gap in training time and structural complexity is significant. Modern research, therefore, focuses on bridging this gap: creating models that approach the accuracy of ensembles while retaining the structural simplicity of single trees.

2.4. Interpretability Evaluation Metrics

To measure these properties scientifically, the literature employs several quantitative metrics. Let $f(x)$ denote the original complex ensemble and $g(x)$ denote the interpretable surrogate model (or explanation).

- *Fidelity*: Measures how accurately the interpretable explanation mimics the predictions of the original complex ensemble. Formally, for a dataset D containing $|D|$ samples, fidelity is often defined as the accuracy of g with respect to f :

$$Fidelity(f, g) = \frac{1}{|D|} \sum_{x \in D} \mathbb{I}\big(f(x) = g(x)\big) \quad (8)$$

Here, $\mathbb{I}(\cdot)$ represents the indicator function, a mathematical operator that yields a value of 1 if the condition inside the parentheses is true (i.e., both models predict the exact same output for the instance x), and 0 if it is false. Consequently, this formula calculates the exact percentage of instances where the surrogate perfectly replicates the black-box model's behavior. High fidelity ensures that the explanation is a trustworthy representation of the original model [15].

- *Complexity*: Quantifies the cognitive load required for a human to comprehend the generated explanation. For decision trees and rule sets, complexity is typically measured by structural properties such as the total number of nodes, the maximum depth of the tree, or the number of logical conditions. Bassan and Bianchini [16] formalize this concept, arguing that for an explanation to be human-tractable, its structural size C (g) must be polynomially bounded:

C (g) ≤ poly(d) (9) (9)

where d is the dimensionality (number of features) of the input. In practical terms, this means that as the number of features increases, the size of the explanation should only grow at a manageable polynomial rate (e.g., linearly or quadratically), rather than exploding exponentially, which would render it incomprehensible to human users.

- *Stability*: Refers to the robustness and consistency of the explanation when the input data is subjected to minor variations. Let ϵ represent a noise added to the original input x. A highly unstable explanation is one where a tiny change in the input leads to a completely different explanation ($g(x) \neq g(x + \epsilon)$), even though the underlying complex model's prediction remains practically unchanged ($f(x) \approx f(x+\epsilon)$). Such behavior in the explanation mechanism significantly reduces user trust, as highly similar input instances should ideally yield similar logical explanations [17].

 doi:10.20944/preprints202604.2101.v1

Together, those metrics highlight that interpretability is a multi-objective concept rather than a single scalar property. In practice, they are tightly coupled and often conflicting in ensemble settings. Increasing fidelity typically requires larger or more complex explanations, which directly increases structural complexity and cognitive load. Conversely, aggressively simplifying an explanation to improve interpretability may reduce fidelity and fail to capture important interactions learned by the ensemble. This tradeoff is central to the design of interpretable ensemble methods and motivates the diverse spectrum of approaches reviewed later in this paper.

2.5. Why Interpretability Matters

Interpretability is no longer a secondary feature but a requirement in safety-critical or regulated domains. Transparent decision structures facilitate auditing, bias detection, and fairness analysis, while supporting legal compliance such as the "right to explanation" in data-protection regulations. In federated or privacy-sensitive contexts, explainable ensembles show that interpretability and distributed computation can coexist [18]. Moreover, inherently interpretable architectures and modern post-hoc visualization dashboards collectively demonstrate that the trade-off between accuracy and transparency is not absolute but adjustable through design.

3. Related Work

This section reviews existing surveys or overviews that are related to interpretability /explainability of tree-based ensembles. We distinguish prior work into three groups: general XAI surveys/overviews, surveys/overviews related to tree-based models, and technical works including original classification taxonomies.

3.1. General XAI Surveys and Overviews

Broader Explainable AI (XAI) surveys provide the necessary context for locating tree ensembles within the wider machine learning landscape. These studies typically aim to establish universal taxonomies applicable to all model types.

Burkart and Huber [6] provide a foundational survey aiming at classifying explainability methods based on the stage of application. They introduce a taxonomy distinguishing between ante-hoc (intrinsic) interpretability and post-hoc explainability. In their analysis, decision trees are cited as the gold standard for ante-hoc transparency, whereas tree ensembles are strictly categorized as black-box models requiring post-hoc surrogates. However, this binary classification tends to overlook recent hybrid architectures that are complex yet constrained.

Similarly, Linardatos et al. [7] reviewed a wide range of interpretability methods with the objective of mapping them to specific machine learning tasks. Their taxonomy divides approaches into model-specific and model-agnostic. While they acknowledge the contribution of feature importance metrics for Random Forests, their review largely treats ensembles as candidates for generic surrogate explainers (like LIME or SHAP), rather than exploring ensemble-specific structural simplification.

While those works are essential, they exhibit specific limitations regarding tree ensembles:

1. *Breadth over Depth*: By covering all ML models, they often reduce tree ensemble interpretability to standard Feature Importance, missing specialized techniques like MaxSAT-based induction.
2. *Lack of Structural Nuance*: They typically classify ensembles as purely "post-hoc" problems, ignoring the emerging class of "inherently interpretable" ensemble architectures.

3.2. Surveys and Overviews Related to Tree Ensembles

While general XAI surveys provide a high-level view, a subset of literature allows for a deeper examination by focusing specifically on tree-based ensembles.

 doi:10.20944/preprints202604.2101.v1

Haddouchi and Berrado [3] provide a comprehensive taxonomy specifically for Random Forests. Their primary goal was to map the landscape of interpretability methods applicable to bagging ensembles. They organized the literature into feature importance analysis, rule extraction, and model simplification. Their work highlights a critical trend: while feature attribution remains the dominant industry standard, there is a growing shift toward rule-based alternatives that offer descriptive fidelity. It is rather a survey (i.e., largely based on mere descriptions) than an overview.

Complementing the above theoretical taxonomies, Aria et al. [19] advanced the field by moving from descriptive classification to empirical validation. While they adopted a similar distinction between Internal Processing (e.g., variable importance) and PostHoc approaches (e.g., rule extraction), their primary contribution is a comparative benchmarking of interpretable surrogates on real-world datasets. Unlike purely qualitative surveys, they quantitatively evaluated the trade-off between fidelity and complexity, specifically pitting inTrees against NodeHarvest. Their experimental results demonstrated that inTrees generally yields superior stability and accuracy, leading them to conclude that post-hoc rule extraction offers the optimal balance for reconstructing the decision logic of Random Forests without sacrificing the ensemble's predictive power.

Parallel to this, Gonçalves and Carvalho [4] focused exclusively on Gradient Boosting Decision Trees (GBDTs). Their review classifies interpretability approaches into Feature Importance (e.g., Gain, SHAP), Simplification (e.g., inTrees, RULECOSI+), and Prototypes. A key insight from their work is the alignment of additive feature attribution (specifically SHAP) with the sequential structure of boosting, contrasting it with modelagnostic tools like LIME which do not leverage the ensemble's internal architecture.

Finally, Sepioło and Ligęza [10] offer a critical overview focused explicitly on treebased ensembles in general. They categorize methods into model-agnostic approaches (such as LIME) and model-specific techniques (such as fusing forests into single trees or using iForest). While this work correctly identifies the need for ensemble-specific explanations, it focuses primarily on older reductionist techniques and relies on the traditional "Agnostic vs. Specific" dichotomy.

Despite their depth, the above reviews leave specific gaps that our overview aims to fill:

1. *Fragmented Landscape*: The literature is split between Random Forest surveys [3] and Boosting surveys [4]. There is a lack of unified overviews that organize and compare interpretability and explainability strategies across all existing paradigms on a common categorization framework.
2. *Outdated Taxonomies*: Even general ensemble overviews like [10] fail to capture the new wave of "Interpretability-by-Design" architectures (e.g., FIGS) or the distinction between extraction and design.
3. *No domain consideration*: Most tree-specific surveys treat interpretability as a purely technical problem, largely overlooking the impact of application domain on the models.

Our work fills in the above gaps in a comprehensive way.

3.3. Technical Works Including Original Classification Taxonomies

Hatwell et al. [15], in proposing the CHIRPS framework, formalize a critical distinction in the rule extraction landscape. They categorize methods into decompositional approaches, which "unpack" the internal representation of the ensemble (e.g., extracting decision paths from every tree), and didactic (or pedagogical) approaches, which treat the model as a black box and learn only from its input-output behavior. Their work argues that while didactic methods are flexible, they often fail to capture what the model actually computes, whereas decompositional methods can guarantee higher fidelity.

From a theoretical standpoint, Bassan and Bianchini [16] formalize the distinction between interpretable and uninterpretable ensembles not by the heuristic method used, but by rigorous computational complexity bounds. They proved a fundamental separation: ensembles of linear models are computationally intractable to interpret (para-NP-hard), whereas ensembles of decision trees fall into tractable complexity classes (XP or FPT) when specific structural parameters are bounded.

 doi:10.20944/preprints202604.2101.v1

Neto and Paulovich [20], in introducing the ExMatrix tool, classify interpretability visualization methods based on their scalability. They argue that traditional Node-Link diagrams (tree visualizations) fail for random forests due to visual clutter, proposing a shift toward Matrix-based metaphors that can handle the massive rule sets generated by bagging ensembles.

Gulowaty and Wozniak [21] synthesize existing literature into a 3D classification framework based on: (i) Timing (Post-hoc vs. Intrinsic), (ii) Specificity (Model-specific vs. Model-agnostic), and (iii) Scope (Global vs. Local). They explicitly position extraction algorithms as "Global, Post-hoc, and Model-Specific," effectively carving out a niche for model distillation techniques.

Lal et al. [22] categorize the field based on the format of the explanation: Rule-based (lists of conditions), Tree-based (single surrogate trees), and Instance-level (local weights). Crucially, they identified a gap in "Minority Class Fidelity," arguing that existing taxonomies often overlook the specific challenge of explaining rare events (e.g., fraud) where global fidelity metrics are misleading.

3.4. Position of the Present Work

The landscape of interpretability research for decision tree ensembles is rich but fragmented. As analyzed in the previous subsections, existing reviews tend to fall into one of three categories, each with distinct limitations:

1. General XAI Surveys (e.g., [6,7]) offer breadth but lack the technical depth required to navigate the specific structural constraints of ensemble learning.
2. Model-Specific Overviews effectively categorize methods for Random Forests [3] or Gradient Boosting [4] in isolation but rarely compare strategies across these paradigms. Crucially, broad decision tree surveys like [23] often explicitly exclude ensembles to maintain focus.
3. Technical Taxonomies (e.g., [15,16]) provide rigorous definitions for specific niches, such as rule extraction fidelity or computational complexity, but do not attempt to map the broader methodological landscape.

Our work differentiates itself by bridging the above gaps. Instead of focusing on a single algorithm class or interpretability technique, we provide a more holistic overview of the interpretability/explainability models published in the period 2020–2025. We extend the fragmented taxonomies of Haddouchi [3] and Gonçalves [4] by integrating them into a unified framework that treats post-hoc and inherent approaches not as afterthoughts, but as primary design pillars. Crucially, this framework is not merely an incremental reorganization of existing literature. By intersecting methodological mechanisms with the interpretability scope, our dual-view classification exposes the fundamental algorithmic trade-offs, such as fidelity versus complexity, that govern tree-based ensembles. This synthesis provides a prescriptive guide, moving the field from ad-hoc tool selection toward systematic, domain-aware interpretable design [26]. By synthesizing algorithmic innovations with the formal complexity bounds identified by Bassan and Bianchini [16], we aim to move the field from a collection of ad-hoc explanation tools toward a structured science of interpretable ensemble design.

Despite the width of the aforementioned surveys, none provides a unified and structured overview that simultaneously (i) spans the full spectrum of decision tree ensemble architectures, (ii) systematically distinguishes between ante-hoc and post-hoc interpretability mechanisms, and (iii) integrates methodological advances with domain-specific deployment considerations. Existing works either focus on a single ensemble family, emphasize explanation techniques without architectural context, or treat interpretability as a secondary property. In contrast, the present overview aims to synthesize these perspectives into a coherent framework that highlights both theoretical advances and practical design trade-offs in modern interpretable tree-based ensembles. Overall, these approaches illustrate the diversity of strategies adopted to combine predictive performance with human interpretability in ensemble settings.

4. Methodology

 doi:10.20944/preprints202604.2101.v1

This study adopts a structured overview methodology designed to integrate, compare, and synthesize interpretability techniques for modern decision-tree ensembles. Our process combines systematic search procedures with qualitative thematic analysis.

4.1. Search Strategies and Sources

We conducted a keyword-based search across academic databases including Google Scholar, arXiv, Scopus, and IEEE Xplore, covering mostly the period 2020–2025. We also have used some older, but fundamental studies. Search terms included combinations of "interpretable ensemble trees", "explainable boosting", "explainable Random Forest", and "visualization of tree ensembles". The search was complemented by backward and forward snowballing from established surveys (e.g., [3,15]) to ensure coverage of both foundational and emergent studies. A total of 62 works were selected for final review.

4.2. Inclusion and Exclusion Criteria

For our main study, concerning categorizations of interpretability methods, papers were selected based on the following criteria:

- *Relevance*: The paper proposed a new interpretability/explainability method specific to decision tree ensembles (Random Forests, GBMs, or hybrid architectures).
- *Methodological clarity*: The interpretability/explainability contribution was explicitly defined (e.g., rule extraction, structural simplification, or visual analytics).

We excluded papers focused solely on predictive performance without interpretability discussion, as well as generic XAI frameworks that did not address the specific constraints of tree-based models.

On the other hand, for our secondary study on surveying methods per application domain, we included papers not only presenting a new interpretability but also applying existing methods on a specific problem domain.

4.3. Data Extraction and Synthesis

Each selected paper was processed through a unified extraction protocol capturing the ensemble model type, the stage of intervention and the interpretability objective. Data synthesis was performed iteratively using a thematic grouping approach. Initially, papers were clustered by their primary technical contribution (e.g., rule extraction, surrogate modeling, or architectural constraints).

These clusters were then refined into the two primary taxonomies presented in this overview: (i) Type of model, dividing the literature into *acquired* (post-hoc) extraction approaches and *inherent* architectures, and (ii) Interpretability scope, distinguishing between model level (global) and instance or feature-level (local) approaches. This qualitative thematic synthesis allowed us to identify cross-cutting trends, such as the adaptation of specific interpretability strategies to rigorous domain constraints (e.g., privacy-preserving federated learning and physical laws), rather than treating them merely as generic algorithms.

5. Categorization of Tree Ensemble Interpretability Models

Given the diversity of interpretability and explainability mechanisms proposed for TBEs, a structured categorization is essential to avoid fragmented or ad-hoc comparisons. Existing literature often confuses the method of interpretation with the scope of the explanation. Prior surveys have adopted varying classification strategies: Gonçalves and Carvalho [4] classify approaches based on specific techniques specifically within Gradient Boosting, while Costa and Pedreira [23] organize developments by broad learning goals.

We propose a more detailed, multi-view categorization that extends these frameworks. We classify the reviewed works through two distinct taxonomies: (i) by methodological mechanism (the

 doi:10.20944/preprints202604.2101.v1

algorithmic approach), and (ii) by interpretability scope (the granularity of interpretability/explainability).

5.1. Based on the Type of the Model

The primary categorization organizes the literature into two fundamental families based on when and how transparency is achieved. It reflects both the stage at which interpretability is introduced and the nature of the explanatory artifact produced. As illustrated in Figure 4, interpretability models are divided into *acquired interpretability/explainability models*, which extract explanations from an already trained black-box model, and *inherent interpretability/explainability models*, which constrain the architecture during training to guarantee transparency by design. The term "acquired" is used as opposed to "inherent"; In this context, traditional term *post-hoc* is used as synonymous to *acquired* in the sequel. Also, the term *ante-hoc* is used as synonymous to *inherent*.

5.1.1. Acquired Interpretability/Explainability Models

Acquired interpretability/explainability models operate on a completely trained, opaque ensemble model. Their primary objective is to reverse-engineer, approximate, or visually represent the decision-making logic without altering the original model's training process or predictive performance.

- *Extracted Rules*: These methods algorithmically distill the dense, overlapping decision paths of a forest into a smaller, symbolic set of logical rules (e.g., IF-THEN statements) that explain the global boundary. The goal is to maximize logical coverage while minimizing the number of rules to prevent cognitive overload [15,22,24–27].
- *Surrogate Models*: This approach frames interpretability/explainability as a "teacher-student" distillation task, involving the training of a secondary, inherently interpretable model to mimic the input-output behavior of the complex ensemble. We distinguish two subcategories:
    - *Tree Surrogates*: Approximating the ensemble using a single, shallow decision tree that balances fidelity with depth [28].
    - *Non-tree Surrogates*: Approximating the ensemble using linear frameworks or alternative transparent structures, such as functional ANOVA models, to capture main effects and interactions cleanly [29,30].
- *Regularization Based*: Rather than extracting new structures, these techniques apply mathematical constraints or smoothing operations to the existing tree structures post-training. This simplifies the decision boundaries by reducing the influence of noisy, deep splits while retaining the overall ensemble architecture [17,31].
- *Modular*: These methods decompose the feature space post-training and apply distinct, localized explanations to specific sub-regions, creating a mosaic of simple explainers. We distinguish two types:
    - *Tree Modular: Using local decision trees dedicated to specific, non-overlapping data regions [21].*
    - *Non-tree Modular: Using alternative local estimators to explain highly specific geometric sub-spaces.*
- *Visualization Oriented*: When symbolic logic or rule sets become too massive to read, these methods translate the ensemble's structural topology or prediction behavior into human-readable visual abstractions. We discern three types of visualization:
    - *Rule-based: Visualizing the topology and overlap of extracted rules in matrix formats [20].*
    - *Feature-based: Visualizing how features contribute to predictions globally or locally via heatmaps or attribution scores [29,32,33].*
    - *Point-based: Explaining the model through the visualization of representative data prototypes [34].*

 doi:10.20944/preprints202604.2101.v1

- *Counterfactuals*: Methods that explain a specific prediction by identifying the minimum necessary changes to an input feature vector required to alter the model's output, offering actionable recourse options. Again, we distinguish three types of counterfactual approaches:
    - *Feature-based: Generating explainability by computationally modifying individual feature values via mathematical optimization to cross the decision boundary [37,39,40,42].*
    - *Set-based: Defining a continuous region or a definitive set of counterfactual conditions, offering robustness guarantees rather than a single point [35,38,41,43].*
    - *Instance-based: Providing explainability based on actual differences between real training instances that lead to different model predictions, grounding the recourse in historical data [36].*

5.1.2. Inherent Interpretability Models

Inherent approaches embed transparency directly into the model's design. By constraining the learning algorithm, these methods ensure that the resulting ensemble is naturally comprehensible to humans without requiring secondary explanation tools.

- *Aggregational*: These methods restrict the way trees are combined. Instead of allowing complex, high-order feature interactions, they typically enforce structural additivity or partial consolidation, allowing the model to be interpreted as a sum of simple, independent functions [13,44,45].
- *Modular*: These architectures dynamically partition the dataset during the training phase itself, explicitly assigning specialized, simple models to distinct regions of the feature space.
    - *Tree Modular*: Using single decision trees as the local experts [46].
    - *Non-tree Modular*: Using alternative interpretable models as local experts [47,48].
- *Distributional*: These frameworks focus on explaining models while strictly managing the underlying data distribution, often integrating interpretability with privacy preserving or federated learning protocols where feature transparency must be balanced with data security [12].

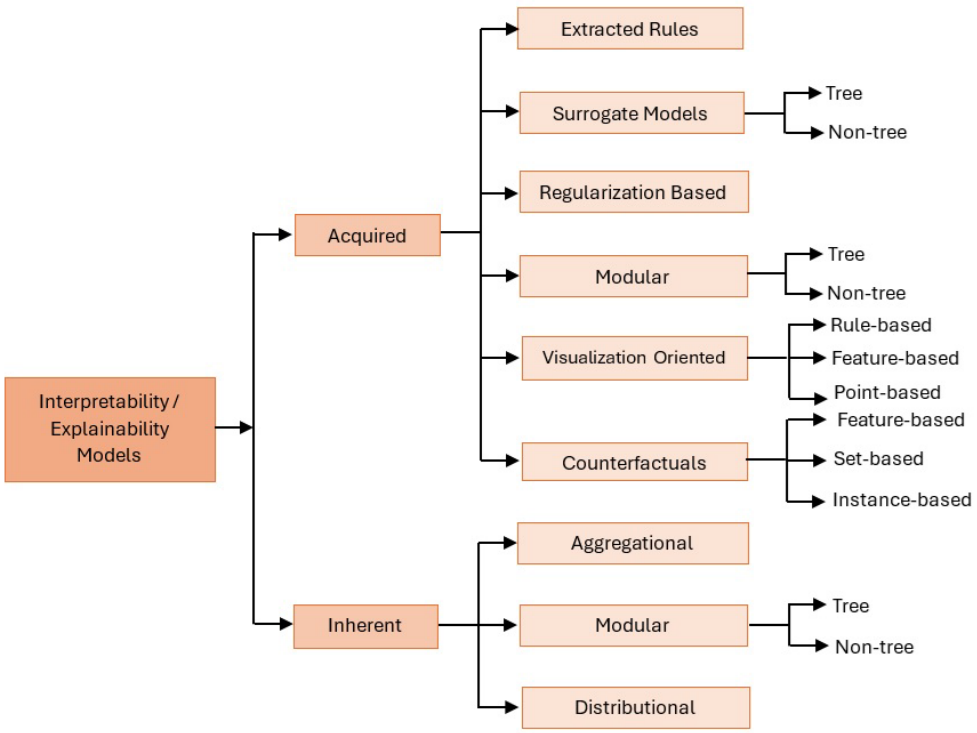

Figure 4. Categorization of interpretability and explainability models based on the type of the model.

To facilitate the practical navigation of this framework, we provide an open-source Interactive Taxonomy Explorer as supplementary material, freely available at:

 doi:10.20944/preprints202604.2101.v1

https://alexandrosmiteloudis.github.io/tree-ensemble-taxonomy/. This interactive tool maps directly to the proposed classification in Figure 4 and explicitly connects each interpretability mechanism to its primary application domains. By integrating these domain-specific actions, such as utilizing causal estimators for medical hypothesis generation, fast local surrogates for engineering risk assessment, or exact rule extraction for legal compliance, the explorer serves as a practical decision-support dashboard for practitioners to select the optimal methodology based on their specific operational constraints.

5.2. Based on the Interpretability Scope

Complementary to the methodological mechanisms (Acquired vs. Inherent), we also categorize the reviewed approaches by their interpretability scope. As shown in Figure 5, this taxonomy defines the scope and granularity of the explanation provided by the method.

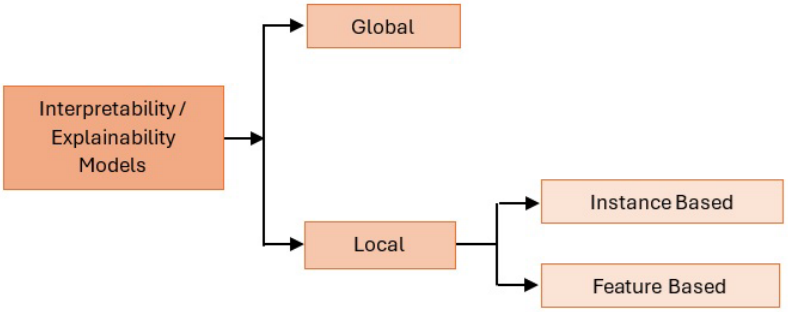

Figure 5. Categorization of interpretability and explainability models based on the scope.

- *Global (Model-Level)*: The goal is to understand the entire model's logic at a macro level, providing a complete overview of how the ensemble makes decisions across the whole feature space. Inherent architectures like FIGS [13] naturally fall here, as their constrained structure allows the user to view the whole model at once. Similarly, global rule extraction methods [15,25] and global surrogate trees target this objective by summarizing the entire forest's boundaries.
- *Local (Instance-Level)*: The goal is to explain the model's behavior for a specific, restricted region or an individual data point. This is critical in high-stakes domains (e.g., healthcare or mining safety) where operators need to know exactly why a specific decision or alarm was triggered. This objective is further divided into:
    - *Instance Based*: Methods that explain a specific prediction for a single data point. Examples include providing counterfactuals (e.g., finding the minimal changes to flip a specific prediction [49]) or utilizing local surrogate approximations.
    - *Feature Based*: Methods that identify which specific variables drove the prediction for a localized subset or outcome. This includes local feature attribution techniques and causal inference frameworks, such as the Causal Rule Ensemble (CRE) [26], which discover subgroups with heterogeneous treatment effects to distinguish true causal drivers from mere correlations in specific instances.

6. Analysis of Algorithmic Mechanisms

Building upon the model-based taxonomy introduced in the previous section, this analysis examines interpretability approaches not as standalone techniques but as design choices situated along multiple trade-off axes. In particular, we analyze how different methods balance fidelity against structural simplicity, global transparency against local explanation, and architectural transparency against post-hoc reconstruction. This perspective enables a comparative evaluation that emphasizes underlying principles and limitations rather than implementation-specific details.

 doi:10.20944/preprints202604.2101.v1

This section critically analyzes the methodological families defined in our taxonomy. We examine the specific algorithmic mechanisms employed by the reviewed studies, evaluating how they extract, or design interpretability based on their structural characteristics.

6.1. Acquired Interpretability/Explainability Models

Methods in this category navigate the fidelity-complexity trade-off by operating on already trained, opaque ensemble models. Their goal is to maximize the logical explanation of the ensemble while keeping the extracted artifacts comprehensible to humans.

6.1.1. Extracted Rules

These methods treat rule extraction as a distillation process to create a symbolic set of conditions. To handle large-scale forests, mining frameworks exploit statistical recurrence. CHIRPS [15] extracts decision paths from each tree that contribute to the majority classification and employs Frequent Pattern (FP) mining to filter the most commonly occurring split conditions, greedily building a concise, high-precision classification rule. Similarly, TE2Rules [22] uses the Apriori algorithm on tree nodes to generate a configurable list of rules, specifically excelling at capturing the logic of the minority class.

Other frameworks focus on refinement and optimal selection. The inTrees framework [24] harvests all root-to-leaf paths, prunes irrelevant variable-value pairs via feature selection, and summarizes them into a Simplified Tree Ensemble Learner (STEL). Pushing for mathematical optimality, Bonasera and Carrizosa [25] formulate rule extraction as a set partitioning problem solved via Integer Programming, extracting an optimal subset that maximizes stability while minimizing loss. In a symbolic parallel, Takemura and Inoue [27] apply Answer Set Programming (ASP) by encoding the voting logic into clauses and utilizing SAT solvers to deduce the minimal set of non-contradictory rules. Finally, the Causal Rule Ensemble (CRE) [26] algorithm shifts the focus to causal inference, generating rules from Random Forests and GBMs and filtering them to estimate Conditional Average Treatment Effects (CATE), thereby uncovering heterogeneous treatment effects rather than simple correlations. This transition highlights a broader trend toward causal interpretability. Beyond acquired rule extraction, methods like Causal Forests adapt the standard decision tree splitting criteria to directly maximize the heterogeneity of treatment effects. This transforms the ensemble from a purely predictive, associational black-box into a transparent causal estimator, which is critical for answering “what-if” intervention questions in applied sciences.

6.1.2. Surrogate Models

Surrogate modeling frames interpretability and explainability as a secondary learning task where a new, transparent model mimics the black box. For tree-based surrogates, Khalifa et al. [28] proposed the Forest-Based Tree (FBT) algorithm, which transforms a Random Forest into a self-explainable decision tree through pruning and conjunction set generation. For non-tree surrogates, Yang et al. [30] leverage the functional ANOVA framework. They transform a fitted tree ensemble into a generalized additive model (GAM) with interaction terms, allowing main effects and pairwise interactions to be precisely interpreted without approximation, aided by an effect purification algorithm.

6.1.3. Regularization Based

Regularization techniques modify the influence of existing structures without fully discarding them. Hierarchical Shrinkage (HS) [31] applies post-hoc regularization by shrinking the prediction over each leaf towards the sample means of its ancestors. This smooths the fragmented decision boundaries of random forests and stabilizes post-hoc interpretability outputs. Inspired by this, Pfeifer et al. [17] proposed a Tree Smoothing technique that adjusts the probabilities at the leaf nodes by

 doi:10.20944/preprints202604.2101.v1

giving more weight to nodes near the root, thus reducing the impact of noisy, deep splits while retaining the overall tree structure.

6.1.4. Modular

Modular post-hoc methods decompose the complex model into localized explanation spaces. The NOTE (Non-Overlapping Tree Ensemble) framework [21] partitions the feature space into disjoint "competence regions" based on induced rules. For each region, a separate local decision tree is trained to mimic the rule's output, creating an understandable modular structure. Taking a non-tree, exact geometrical approach, Blanchart [49] computes the intersections of leaf nodes across all trees to decompose the feature space into disjoint multidimensional structures, allowing for the direct, analytical calculation of optimal counterfactuals.

6.1.5. Visualization Oriented

When logic is too dense, visual abstractions serve as proxies. The Ex-Matrix framework [20] abandons the classic node-link tree diagram for a rule-based matrix visualization where rows represent extracted disjoint rules, columns represent features, and cells represent rule predicates. For feature-based visualization, Di Teodoro et al. [29] introduced VITE, a hierarchical tool that visualizes feature usage frequency across tree levels via heatmaps, assuming features closer to the root are the most significant. Furthermore, to address privacy constraints, Jetchev et al. [33] developed XorSHAP, a privacy-preserving algorithm that computes SHAP values via secure multiparty computation (SMPC), revealing feature importance without exposing raw data. Furthermore, the computation of these attributions has been significantly accelerated by advanced SHAP variants [4]. While generic SHAP is computationally prohibitive for large ensembles, algorithms like TreeSHAP leverage the internal structure of trees to compute exact Shapley values in polynomial time, a property increasingly utilized in modern surveys [4,33]. Building on this foundation, recent state-of-the-art applications have adapted these variants to handle highly correlated features, ensuring that feature attributions in dense forests do not assign undue importance to dependent variables in complex domains like medicine and epidemiology [51,53].

Expanding on feature attribution, Chen et al. [32] proposed a hybrid ensemble blending method in Quantitative Structure-Property Relationship (QSPR) modeling, which combines the feature importance rankings of multiple distinct tree ensembles into a single, unified interpretability score. Finally, moving to point-based explanations, Tree Space Prototypes (TSP) [34] explain predictions by identifying representative training points (prototypes) using distance metrics derived directly from the tree ensemble's internal geometry and a modified k-medoids algorithm.

6.1.6. Counterfactuals

Counterfactuals explain a prediction by describing the minimal changes required to flip the outcome. Generating these for non-differentiable trees is mathematically challenging. Based on our taxonomy, the reviewed methods approach this problem through three distinct mechanisms:

Feature-based Counterfactuals: These approaches utilize mathematical optimization to directly modify input features. Optimization-based frameworks like FOCUS [37] solve the non-differentiability issue by smoothing discrete split conditions into continuous approximations (e.g., sigmoid functions), allowing gradient-descent optimization to find exact counterfactuals. Similarly, OCEAN [42] tackles optimization via Mixed-Integer Programming (MIP), using isolation forests to enforce plausibility constraints during feature modification. To guarantee actionable real-world recourse, DiCE [40] minimizes distance metrics while strictly preserving medical constraints for specific features, and RobX [39] ensures that the feature perturbations provided as explanations remain robust even against future model shifts.

Set-based Counterfactuals: Instead of outputting a single modified data point, these methods define entire regions or sets of conditions where the flipped outcome is guaranteed. FACET [38]

 doi:10.20944/preprints202604.2101.v1

computes the intersection of leaf regions to define an n-dimensional volume (a counterfactual region), offering users a minimum robustness guarantee regarding how much a feature can deviate safely. RF-OCSE [35] defines optimal counterfactual sets by performing a partial fusion of the Random Forest predictors into a single tree. Further expanding on spatial division, the Counterfactual Explanation Tree (CET) [41] constructs an explicitly interpretable decision tree via stochastic local search to partition the entire input space, assigning an optimal set of counterfactual actions to each distinct sub-region.

Instance-based Counterfactuals: These methods ground their explanations in actual data samples rather than artificial perturbations. Harvey et al. [36] propose a model-aware framework that uses the internal geometry of the Random Forest itself (specifically leaf co-occurrences) to identify real, historically similar training instances as counterfactuals, tracing the decision boundary between them. Operating on a similar localized logic, LORE [43] uses genetic algorithms to build a synthetic local neighborhood around a specific instance, training a local decision tree to extract factual and counterfactual rules grounded directly in the neighborhood's data distribution.

6.2. Inherent Interpretability Models

This category shifts the focus from post-hoc approximation to ante-hoc design, constraining the hypothesis space during training to ensure transparency without the need for secondary explainer tools.

6.2.1. Aggregational

The opacity of standard boosting arises from high-order feature interactions entangled across the ensemble. Fast Interpretable Greedy-tree Sums (FIGS) [13] resolves this by enforcing structural additivity. The algorithm grows a flexible number of trees simultaneously but restricts the total number of splits. The ensemble prediction becomes a sum of independent functions.

This decomposition collapses the interaction complexity, allowing users to inspect the contribution of each feature subset Sj in isolation. Following a similar additive logic, Konstantinov and Utkin [44] construct multiple simple GBMs in parallel, where each tree uses only a single feature with a depth of 1, combining them via a Lasso-optimized weighted sum. Attempting a hybrid approach, PCTBagging [45] builds a single "consolidated" tree for the top decision levels to provide a unified global logic, and branches into bagging ensembles only at the leaves to reduce variance.

6.2.2. Modular

Modular architectures dynamically route instances to specialized sub-models. The Mixture of Decision Trees (MoDT) [46] relies on a gating function to route a prediction through a small, single-digit number of decision trees. Similarly, the League of Experts (LoE) [47] explicitly partitions the feature space into disjoint subsets and assigns one specialized, simple ensemble member (an expert) to each subset. Pushing for bottomup simplicity, Arwade and Olafsson [48] identify subsets with locally significant linearly separable classes in an iterative, clustering-like manner, assigning a simple decision tree to each structure.

6.2.3. Distributional

In decentralized and privacy-preserving environments, inherent interpretability must be maintained without exposing underlying feature values. In Vertical Federated Learning, where features are split across different data owners, the Fed-EINI framework [12] secures the decision paths using additively homomorphic encryption. This allows the model to inherently disclose the meanings of the Host parties' features to the User party securely, maintaining structural interpretability ...without incurring the unaffordable computation overhead associated with fully homomorphic encryption. Expanding on this intersection of privacy and interpretability, the Explainable Federated Learning (XFL) model [50] integrates decentralized training with established

 doi:10.20944/preprints202604.2101.v1

post-hoc techniques like SHAP and LIME. This hybrid approach enables real-time, privacy-preserving predictive maintenance in industrial settings, providing operators with transparent risk factors without compromising decentralized data security.

Explainability in *Federated Learning* (FL) represents a critical intersection of our proposed taxonomy, as FL architectures inherently fragment the tree ensemble across multiple decentralized nodes. To achieve transparency without centralizing data, state-of-the-art federated frameworks must hybridize both inherent and acquired methodologies. During the training phase, they rely on inherent distributional constraints, such as secure multiparty computation or homomorphic encryption, to securely build the ensemble [12,33]. Post-training, they deploy acquired feature-based techniques at the local client level [50]. This cross-taxonomy integration is essential as it ensures that the global ensemble remains structurally secure while still providing local stakeholders with acquired, high-fidelity explanations [12,50].

6.3. Comparative Analysis and Practical Trade-Offs

While the algorithmic mechanisms detailed above offer diverse pathways to interpretability, their practical deployment is governed by strict operational trade-offs. To move beyond descriptive classification, we provide a comparative analysis of the primary methodological families against four critical real-world constraints: Scalability, Computational Cost, Robustness/Stability, and Real-World Usability. The following assessments represent a qualitative synthesis derived from the theoretical complexity bounds and empirical observations reported in the reviewed literature, concisely depicted in Table 1.

- **Extracted Rules (Heuristic):**
  - *Trade-offs:* These methods offer high scalability and low-to-moderate computational cost, as they rely on fast pattern mining [15,22]. However, they exhibit moderate stability, as heuristic extraction is prone to variance when training data is noisy.
  - *Usability:* High, provided the extracted rule lists are kept short to avoid human cognitive overload.
- **Extracted Rules (Exact/MILP):**
  - *Trade-offs: By relying on mathematical optimization, these methods offer high robustness and absolute logical certainty [25,27]. The critical trade-off is their very low scalability and extremely high computational cost, as the underlying search spaces are often NP-hard for large ensembles [16,25].*
  - *Usability: Very high for compliance-heavy domains requiring absolute guarantees.*
- **Surrogate Models:**
  - *Trade-offs: Surrogates excel in scalability and have low computational costs due to their model-agnostic training [28,30]. However, they suffer in stability and robustness due to the "fidelity gap"—the inherent approximation error between the surrogate and the true black-box boundary.*
  - *Usability: High, as they output familiar formats like single trees or linear equations.*
- **Counterfactuals:**
  - *Trade-offs: These approaches have lower scalability and high computational costs because they require iterative optimization or mixed-integer programming for specific queries [37,42]. However, set-based counterfactuals offer high robustness guarantees [38].*
  - *Usability: Very high, as they provide immediately actionable recourse and "what-if" scenarios for end-users.*
- **Inherent Architectures (Aggregational & Modular):**
  - *Trade-offs: Inherent models impose structural constraints during the training phase, leading to moderate computational costs during induction compared to standard bagging [13,46,47]. However, they offer high scalability at inference and high robustness, as their additive or routed structures naturally reduce variance [44].*
  - *Usability: High, providing immediate global or localized transparency without the need for secondary, error-prone explanation tools.*

 doi:10.20944/preprints202604.2101.v1

Table 1. Mapping of Methodological Families to Critical Real-World Constraints.

Methodological Family	*Computational Cost*	*Scalability*	*Robustness & Stability*	*Real-Word Usability*
Extracted Rules (Heuristic)	Low to Moderate	High	Low to Moderate (Prone to instability with noisy data)	High
Extracted Rules (Exact/MILP)	Very High (NP-hard search space)	Very Low	High (Mathematically guaranteed optimal boundaries)	High
Surrogate Models	Low	High	Moderate (Limited by the fidelity gap between the surrogate and the black box)	High
Counterfactuals	High (Requires iterative optimization or gradient descent per query)	Low	Moderate to High (Set-based methods offer strict robustness guarantees)	Very High
Inherent (Aggregational)	Moderate (Higher than standard bagging due to structural constraints)	Moderate	High (Additive structures inherently reduce variance and noise)	High
Inherent (Modular)	Moderate	High	High (Local experts remain stable within their defined feature regions)	High

7. Domain-Specific Methodological Trends

The adoption of interpretable ensembles is not universal across industries. Our analysis reveals that the choice of methodology (as defined in our taxonomy in Section 5) is heavily dictated by the specific regulatory and operational constraints of the application domain. This section maps these sectoral preferences, synthesizing the trends observed in the technical analysis.

7.1. Healthcare

In medical diagnostics, the primary constraint is biological plausibility. Clinicians require explanations that align with known physiological pathways to trust a prediction. Consequently, this sector favors Causal Attribution and Feature Interaction analysis over simple rule extraction. Haque et al. [51] utilize a stacked ensemble for liver disease detection, explicitly employing SHAP-based attribution to validate that the model's ranking of biomarkers (e.g., total bilirubin) matches medical literature. Similarly, in survival analysis for heart failure, Moreno-Sanchez [52] prioritizes interaction detection. The goal is not just to predict mortality but to identify non-linear risk factors (e.g., how age modulates the risk of high blood pressure), a task for which standard additive models are insufficient without explicit interaction analysis.

In the sector of public health and epidemiology, interpretability is equally vital for understanding disease transmission dynamics and formulating intervention policies. Zheng et al. [53] demonstrated this by developing a data-driven ensemble framework (combining Random Forest and XGBoost) to forecast the occurrence of COVID-19. Rather than proposing a new interpretability extraction mechanism, they leveraged SHAP values to conduct both global and local feature dependence analyses. Their application of these post-hoc tools allowed epidemiologists to quantify the non-linear impacts and critical thresholds of specific drivers, such as population flow,

 doi:10.20944/preprints202604.2101.v1

temperature, and air quality (PM2.5), providing actionable, transparent insights for localized disease control without sacrificing the high predictive accuracy of the underlying boosting and bagging models.

From a practical standpoint, while acquired (post-hoc) methods like SHAP provide scalable feature attribution for such high-dimensional data [51,53], practitioners requiring actionable clinical interventions should prioritize causal architectures or exact rule extraction to isolate true physiological drivers from spurious correlations.

7.2. Industrial and Physical Engineering

In civil and environmental engineering, models must respect fundamental physical laws. Interpretability here serves as a debugging tool to ensure the model has not learned spurious correlations that violate physics. Jia et al. [54] apply ensemble interpretation to concrete compressive strength. They utilize global feature dependence plots to confirm that the relationship between water-cement ratio and strength follows the expected inverse law, determining that the model is safe for structural design. In environmental modeling, Song et al. [55] (solar radiation) and Li et al. [56] (water quality) rely on Global Feature Relevance. The operational requirement is to identify the physical drivers of environmental change (e.g., rainfall vs. pollution sources) to guide regulatory intervention.

In safety-critical underground engineering, such as mining and tunneling, interpretability is crucial for real-time risk assessment and disaster prevention. Qiu and Zhou [57] developed a stacking ensemble learning model to predict short-term rockbursts utilizing microseismic monitoring data. They integrated established post-hoc techniques, namely SHAP, LIME, Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE), to extract both global and local explanations. From a global perspective, SHAP and PDP allowed researchers to validate the model against physical expectations. More importantly for operational safety, local attribution via LIME and SHAP enabled site operators to understand the specific seismic indicators triggering an individual alarm in real-time, thereby facilitating targeted, timely, and trusted evacuation protocols. This highlights a critical deployment guideline: in real-time, safety-critical environments [57], the high computational overhead of global rule extraction is operationally prohibitive. Practitioners must instead rely on inherently modular architectures or fast, local surrogates (such as LIME) to ensure explanations meet strict latency constraints.

Furthermore, in physical systems driven by sensor data, environments are rarely static. Data distributions frequently experience concept drift, rendering static global explanations obsolete. To address non-stationary time series forecasting, Saadallah [58] deployed an online adaptive local interpretable tree ensemble framework (OLITE). This approach utilizes an interpretable gating tree to dynamically route real-time instances to localized, simple linear models. By coupling this structure with drift detection, the system guarantees that as the time-series pattern shifts, the active sub-ensemble remains both accurate and strictly locally interpretable.

7.3. Finance, Law and Social Sciences

The financial, legal, and social sectors are governed by strict statutory requirements (e.g., GDPR, Equal Credit Opportunity Act). These regulations often mandate transparent decision-making and specific reasons for denying service or making critical social predictions. To address this, Vultureanu-Albisi and Badica [59] employ local surrogates (LIME) to provide student-specific intervention reasons in educational performance prediction. Furthermore, the necessity for legal compliance has driven the development of fairness-aware interpretability frameworks. As highlighted by Arévalo-Cordovilla and Peña [60], these methods actively decompose feature contributions to isolate direct discrimination (e.g., based on gender or race) from proxy discrimination. By tightly coupling interpretability with fairness metrics, auditors can algorithmically detect and rectify biased decision paths within the ensemble's weighting mechanisms, ensuring equitable outcomes for protected demographic groups [60].

 doi:10.20944/preprints202604.2101.v1

Consequently, for practical deployment in heavily regulated environments, heuristic post-hoc approximations carry unacceptable compliance risks due to the fidelity gap. Practitioners should instead mandate exact rule extraction or set-based counterfactuals to provide the absolute logical certainty required for formal legal auditing [60].

7.4. Privacy Preserving Applications

An emerging domain constraint is the strict preservation of data privacy, where interpretability must be achieved without exposing the underlying raw features. In Vertical Federated Learning (VFL), where datasets are distributed across different institutions (e.g., multiple hospitals or banks), frameworks like Fed-EINI [12] employ additively homomorphic encryption to secure the decision paths of tree ensembles, allowing the model to inherently disclose feature importance without breaching data confidentiality. Similarly, Jetchev et al. [33] introduce XorSHAP, which utilizes secure multiparty computation (SMPC) to extract post-hoc SHAP values across decentralized silos. In industrial settings (Industry 4.0), Alshkeili et al. [50] successfully merged these fields by proposing an Explainable Federated Learning (XFL) model that provides technicians with interpretable, real-time predictive maintenance alerts (via SHAP and LIME) without exposing the proprietary sensor data of individual manufacturing plants.

7.5. Summary

Table 2 summarizes this analysis, correlating each domain with its dominant methodological preference and the underlying operational rationale, reflecting the models categorized earlier in type-based taxonomy.

Table 2. Mapping of Application Domains to Preferred Interpretability Models.

Sector	*Dominant Model*	*Operational Rationale*	*Key Studies*
Healthcare	Feature Attribution & Interaction (Post-hoc)	Need for biological plausibility and validation against medical literature.	[51–53]
Engineering	Feature Dependence & Adaptive Modular Ensembles	Validation against physical laws and adaptation to sensor concept drift.	[54–58]
Finance/Social	Local Surrogates & Fairness Constraints	Legal compliance; requirement for auditable instance-level reasons and bias detection.	[59,60]
Privacy Preserving	Cryptographic Extraction & Federated Architectures	Generating explanations while strictly preventing raw data reconstruction across silos.	[12,33,50]

A central insight emerging from this review is that no single interpretability approach can be considered universally optimal for decision tree ensembles. Instead, interpretability is inherently context-dependent, shaped by factors such as data dimensionality, regulatory constraints, user expertise, and deployment environment. Methods emphasizing inherent transparency often sacrifice flexibility or scalability, while post-hoc techniques preserve predictive performance at the cost of approximation and potential instability.

This observation reinforces the importance of aligning interpretability strategies with application-specific requirements rather than pursuing a one-size-fits-all solution. In practice, this often leads to hybrid systems that combine interpretable model structures with selective post-hoc explanations, reflecting a pragmatic balance between transparency, usability, and performance.

8. Practical Decision Framework

 doi:10.20944/preprints202604.2101.v1

To bridge the gap between theoretical taxonomies and real-world deployment, this section proposes a prescriptive decision framework. Selecting the optimal interpretability mechanism is not a one-size-fits-all process; it requires evaluating three intersecting criteria: (a) dataset properties (dimensionality and scale), (b) domain constraints (regulatory compliance and inference latency), and (c) interpretability requirements (global model audits vs. local instance recourse). Based on the trade-offs established in Section 6.3, we outline four primary decision pathways:

- **High-Stakes, Regulated Domains** (e.g., Finance, Law):
  - *Requirements*: Absolute logical certainty for formal auditing; global or regional scope.
  - *Recommendation*: Exact Rule Extraction (via MILP) or Set-based Counterfactuals [25,38].
  - *Trade-off*: These methods guarantee maximum fidelity and mathematical robustness, but their extreme computational complexity strictly limits their scalability to smaller datasets and shallower ensembles.
- **Safety-Critical, Real-Time Systems** (e.g., Engineering, Sensor Networks):
  - *Requirements: Ultra-low inference latency; ability to handle concept drift; local scope.*
  - *Recommendation: Inherent modular architectures or fast, local instance-based surrogates (e.g., LIME) [57,58].*
  - *Trade-off*: This approach maximizes operational scalability and real-time usability for immediate intervention, but inherently sacrifices global structural transparency.
- **Exploratory and Scientific Domains** (e.g., Healthcare, Epidemiology):
  - *Requirements: High-dimensional dataset handling (e.g., genomics, biomarkers); identification of true physiological drivers.*
  - *Recommendation: Advanced post-hoc feature attribution (e.g., TreeSHAP) or Causal Rule Ensembles [26,53].*
  - *Trade-off: Offers excellent scalability for massive feature spaces. However, practitioners must accept a potential "fidelity gap" where post-hoc attributions might misrepresent highly correlated features, requiring causal architectures for true intervention analysis.*
- **Privacy-Preserving Environments** (e.g., Decentralized Silos):
  - *Requirements: Strict data confidentiality; distributed training across multiple institutions.*
  - *Recommendation: Distributional inherent models utilizing cryptographic protocols (e.g., Federated SHAP) [12,50].*
  - *Trade-off: Ensures strict data privacy and structural security, but incurs a massive computational overhead during explanation generation, rendering it unsuitable for low-latency applications.*

9. Discussion and Open Challenges

The methodological landscape of interpretable and explainable TBEs is currently undergoing a fundamental bifurcation. Our analysis indicates a divergence between reconstructive approaches, which attempt to distill logic from opaque forests, and constructive approaches, which constrain the ensemble's architecture ab initio.

The reconstructive paradigm, dominated by rule extraction, faces a severe computational ceiling. While optimization-based methods like those proposed by Bonasera and Carrizosa [25] offer mathematical guarantees of fidelity via MILP, their scalability is inversely proportional to the ensemble size. For a Random Forest with hundreds of trees, the search space for a globally optimal subset of non-conflicting rules becomes intractably large. Consequently, practitioners in high-throughput environments, such as real-time financial monitoring, are forced to rely on heuristic mining frameworks like TE2Rules [22]. These mining approaches trade exactness for speed, leveraging statistical recurrence rather than logical proof. While efficient, this approximation introduces a "fidelity gap," where the extracted explanation may diverge from the model's actual

 doi:10.20944/preprints202604.2101.v1

decision boundary in edge cases, a risk that remains unacceptable in safety-critical engineering domains [57].

In response to these limitations, the field is shifting toward inherently interpretable designs. The FIGS framework [13] demonstrates that high predictive accuracy does not strictly require opaque complexity. By replacing the greedy, high-variance induction of standard bagging with evolutionary optimization or structural additivity, these methods achieve competitive performance with significantly simpler logical structures. This trend suggests that the historical trade-off between accuracy and interpretability is not an immutable law, but rather a consequence of legacy algorithms (like standard Boosting) that were not designed with transparency as a primary objective. Importantly, these tradeoffs are not static and often depend on the application context, data characteristics, and regulatory requirements.

However, a critical unresolved challenge remains the stability of interpretations. Decision trees are notoriously sensitive to small perturbations in training data. As noted by Pfeifer et al. [17], this instability propagates to feature attribution scores, leading to scenarios where two statistically similar instances receive contradictory explanations. In dynamic environments, ensuring that explanations remain temporally consistent while the model adapts to concept drift is paramount. Without this stability, end-users, whether clinicians or loan officers, cannot develop trust in the system, regardless of the explanation's theoretical fidelity.

Additionally, the integration of interpretability and explainability with privacy-preserving distributed systems introduces a new frontier of challenges. While cryptographic protocols like homomorphic encryption and secure multiparty computation enable transparent federated ensembles without exposing raw data [12,33], balancing the massive computational overhead of these cryptographic explanations with real-time operational needs remains an open problem.

To advance the field from theoretical proposals to reliable deployments, future research must address four critical open challenges:

- **Standardized "Actionability" Metrics:** The field currently optimizes for sparsity or surrogate fidelity, which often fail to correlate with human cognitive fit. A rule set might be mathematically minimal yet practically unactionable if it relies on immutable features. Future work must develop standardized, human-grounded evaluation protocols that penalize domain irrelevance and measure the real-world actionability of the generated explanations.
- **Temporal Stability Under Concept Drift:** Decision trees are notoriously sensitive to data perturbations. A major open gap is ensuring that local explanations and counterfactuals remain temporally consistent as the underlying data distribution experiences concept drift in dynamic environments.
- **Scalable Privacy-Preserving Interpretability:** While initial frameworks for explainable federated learning exist, balancing the massive computational overhead of cryptographic explanations (e.g., Secure Multiparty Computation) with real-time operational needs remains a highly restrictive bottleneck that future architectures must resolve.
- **Exploration of "Blind Spots" in the Methodological Taxonomy:** Our proposed classification framework (Figure 4) exposes distinct, under-researched niches within the XAI landscape. Notably, there is a prominent absence of literature concerning *acquired modular non-tree* approaches. While post-hoc modular trees (which partition the space for local decision trees) and inherent non-tree modules are well-documented, extracting regionalized algebraic models from an already-trained ensemble remains a significant gap. Future research should prioritize algorithms capable of this regional non-tree decomposition, bridging the gap between exact local fidelity and global structural understanding.

10. Conclusions

This paper presents an overview of interpretability and explainability approaches for decision TBE models. By synthesizing recent research across rule extraction, inherently interpretable

 doi:10.20944/preprints202604.2101.v1

architectures, visualization tools, and formal analysis, we highlighted how different methodological choices address the transparency limitations introduced by ensemble learning.

Our analysis shows that interpretability and explainability in TBEs is not a binary property, but a spectrum shaped by model structure, explanatory intent, and application context. While acquired interpretability/explainability techniques enable insight into otherwise opaque models, inherently interpretable ensembles demonstrate that transparency can also be embedded directly into model design. Formal and theoretical advances further clarify the structural trade-offs that govern the balance between predictive accuracy and human comprehensibility.

A key conclusion of this overview is that no single interpretability strategy is universally applicable. Instead, effective solutions depend on domain-specific constraints, regulatory requirements, and the expertise of end users. As a result, hybrid approaches that combine transparent model components with targeted explanation mechanisms appear particularly promising.

A critical paradigm shift observed in this review is the transition from algorithm-centric explanations to domain-centric interpretability. Future research directions must prioritize the development of standardized, human-grounded evaluation protocols, the refinement of robust counterfactual generation, and the seamless integration of interpretable ensemble methods into privacy-sensitive and distributed federated learning environments.

By consolidating recent methodological advances under a coherent analytical framework, this overview aims to support both researchers and practitioners in navigating the rapidly evolving field of interpretable machine learning. As decision tree ensembles continue to dominate tabular data tasks and safety-critical applications, advancing their inherent and acquired transparency is no longer a secondary objective, but a fundamental prerequisite for building responsible, accountable, and trustworthy AI systems.

Author Contributions: Conceptualization, A.M.; methodology, A.M. and I.H.; investigation, A.M. and I.H.; visualization, A.M. and I.H.; writing—original draft preparation, A.M.; writing—review and editing, I.H. and A.M.; supervision, I.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: No data used.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Wadsworth: Belmont, CA, USA, **1984**.
2. Quinlan, J.R. C4.5: *Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, USA, **1993**.
3. Haddouchi, M.; Berrado, A. A survey and taxonomy of methods interpreting random forest models. *arXiv* **2024**, arXiv:2407.12759.
4. Gonçalves, V.; de Carvalho, V. A review of interpretability methods for gradient boosting decision trees. *J. Braz. Comput. Soc.* **2025**, 31, 1.
5. Rudin, C.; Kim, B. *Interpretable Machine Learning: Fundamental Principles and Ten Grand Challenges*; Springer: Cham, Switzerland, **2021**.
6. Burkart, N.; Huber, M.F. A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.* **2021**, 70, 245–317.
7. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. *Entropy* **2021**, 23, 1.
8. Minh, D.; Wang, H.X.; Li, Y.F.; Nguyen, T.N. Explainable artificial intelligence: A comprehensive review. *Artif. Intell. Rev.* **2022**, 55, 3503–3568.

 doi:10.20944/preprints202604.2101.v1

9. Nagahisarchoghaei, M.; Nur, N.; Cummins, L.; Nur, N.; Karimi, M.M.; Nandanwar, S.; Bhattacharyya, S.; Rahimi, S. An empirical survey on explainable AI technologies: Recent trends, use-cases, and categories from technical and application perspectives. *Electronics* **2023**, 12, 1092.
10. Sepiolo, D.; Ligeza, A. Towards explainability of tree-based ensemble models: A critical overview. In *New Advances in Dependability of Networks and Systems (Proceedings of the 17th International Conference on Dependability of Computer Systems DepCoS-RELCOMEX)*, Wrocław, Poland, 27 June–1 July 2022; Springer: Cham, Switzerland, 2022; Volume 484, pp. 287–296.
11. Kotsiantis, S.B. Decision trees: A recent overview. *Artif. Intell. Rev.* **2013**, 26, 159–190.
12. Chen, X.; Zhou, S.; Yang, K.; Fao, H.; Wang, H.; Wang, Y. Fed-EINI: An efficient and interpretable inference framework for decision tree ensembles in vertical federated learning. In *Proceedings of the IEEE International Conference on Big Data*, Orlando, FL, USA, 15–18 December 2021; pp. 1242–1248.
13. Tan, Y.S.; Singh, C.; Nassar, K.; Agarwal, A.; Yu, B. Fast interpretable greedy-tree sums (FIGS). *Proc. Natl. Acad. Sci*. USA **2023**, 120, e2218840120.
14. Zharmagambetov, A.; Hada, S.S.; Carreira-Perpinan, M.A.; Gabidolla, M. An experimental comparison of old and new decision tree algorithms. *arXiv* **2020**, arXiv:1911.03054.
15. Hatwell, J.; Gaber, M.M.; Azad, R.M.A. CHIRPS: Explaining random forest classification. *Mach. Learn.* **2024**, 113, 4683–4719.
16. Bassan, I.; Bianchini, M. What makes an ensemble (un)interpretable? A complexity perspective. *arXiv* **2025**, arXiv:2506.08216.
17. Pfeifer, B.; Gevaert, A.; Loecher, M.; Holzinger, A. Tree smoothing: Post-hoc regularization of tree ensembles for interpretable machine learning. *Inf. Sci.* **2024**, 658, 120015.
18. Wang, Z.; Gai, K. Decision tree-based federated learning: A survey. *Blockchains* **2024**, 2, 40–60.
19. Aria, M.; Cuccurullo, C.; Gnasso, A. A comparison among interpretative proposals for random forests. *Mach. Learn. Appl.* **2021**, 6, 100094.
20. Neto, M.P.; Paulovich, F.V. Explainable matrix—visualization for global and local interpretability of random forest classification ensembles. *IEEE Trans. Vis. Comput. Graph*. **2021**, 27, 1427–1437.
21. Gulowaty, B.; Wozniak, M. Extracting interpretable decision tree ensemble from random forest. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China (Virtual), 18–22 July 2021; pp. 1–8.
22. Lal, G.R.; Chen, X.; Mithal, V. TE2Rules: Explaining tree ensembles using rules. *arXiv* **2024**, arXiv:2206.14359.
23. Costa, V.G.; Pedreira, C.E. Recent advances in decision trees: An updated survey. *Artif. Intell. Rev*. **2023**, 56, 4765–4800.
24. Deng, H. Interpreting tree ensembles with inTrees. *Int. J. Data Sci. Anal.* **2018**, 7, 277–287.
25. Bonasera, L.; Carrizosa, E. A unified approach to extract interpretable rules from tree ensembles via integer programming. *arXiv* **2025**, arXiv:2407.00843.
26. Bargagli-Stoffi, F.J.; Cadei, R.; Mock, L.; Lee, K.; Dominici, F. Causal rule ensemble: Interpretable discovery and inference of heterogeneous treatment effects. *arXiv* **2024**, arXiv:2009.09036.
27. Takemura, A.; Inoue, K. Generating explainable rule sets from tree-ensemble learning methods by answer set programming. In *Proceedings of the 37th International Conference on Logic Programming (ICLP)*, Porto, Portugal (Virtual), 20–27 September 2021.
28. Khalifa, F.A.; Abdelkader, H.M.; Elsaid, A.H. An analysis of ensemble pruning methods under the explanation of random forest. *Inf. Syst.* **2024**, 120, 102310.
29. Di Teodoro, G.; Monaci, M.; Palagi, L. Unboxing tree ensembles for interpretability: A hierarchical visualization tool. Eur. *J. Comput. Optim.* **2024**, 12, 100063.
30. Yang, Z.; Sudjianto, A.; Li, X.; Zhang, A. Inherently interpretable tree ensemble learning. *arXiv* **2024**, arXiv:2410.19098.
31. Agarwal, A.; Tan, Y.S.; Ronen, O.; Singh, C.; Yu, B. Hierarchical shrinkage: Improving the accuracy and interpretability of tree-based methods. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, Baltimore, MD, USA, 17–23 July 2022; pp. 111–135.

 doi:10.20944/preprints202604.2101.v1

32. Chen, C.-H.; Tanaka, K.; Kotera, M.; Funatsu, K. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR. J. *Cheminform*. **2020**, 12, 67.
33. Jetchev, D.; Vuille, M. XorSHAP: Privacy-preserving explainable AI for decision tree models. *Cryptology ePrint Archive*, Paper 2023/1859, **2023**.
34. Tan, S.; Soloviev, M.; Hooker, G. Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the 2020 ACM-IMS Foundations of Data Science Conference*, Virtual Event, USA, 19–20 October 2020.
35. Fernández, R.R.; de Diego, I.M.; Aceña, V.; Fernández-Isabel, A.; Moguerza, J.M. Random forest explainability using counterfactual sets. *Inf. Fusion* **2020**, 63, 196–207.
36. Harvey, J.S.; Feng, G.; Zhao, T. Interpretable model-aware counterfactual explanations for random forest. *arXiv* **2025**, arXiv:2510.27397.
37. Lucic, A.; Oosterhuis, H.; Haned, H.; de Rijke, M. FOCUS: Flexible optimizable counterfactual explanations for tree ensembles. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, Virtual Event, 22 February–1 March 2022.
38. VanNostrand, P.M.; Zhang, H.; Hofmann, D.M.; Rundensteiner, E.A. FACET: Robust counterfactual explanation analytics. *Proc. ACM Manag. Data* **2023**, 1, 4.
39. Dutta, S.; Long, J.; Mishra, S.; Tilli, C.; Magazzeni, D. Robust counterfactual explanations for tree-based ensembles. In *Proceedings of the 39th International Conference on Machine Learning (ICML/PMLR)*, Baltimore, MD, USA, 17–23 July 2022.
40. Monson, M.; Sabarmathi, G. From prediction to action: Counterfactual explanations and ensemble learning for explainable maternal health risk modelling. In *Proceedings of the 2nd International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, Bangalore, India, 11–12 July 2025; IEEE: 2025; pp. 1–7.
41. Kanamori, K.; Takagi, T.; Kobayashi, K.; Ike, Y. Counterfactual explanation trees: Transparent and consistent actionable recourse with decision trees. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Virtual Event, 28–30 March 2022.
42. Parmentier, A.; Vidal, T. Optimal counterfactual explanations in tree ensembles. In *Proceedings of the 38th International Conference on Machine Learning (ICML/PMLR)*, Virtual Event, 18–24 July 2021; Volume 139, pp. 8422–8431.
43. Guidotti, R.; Monreale, A.; Ruggieri, S.; Naretto, F.; Turini, F.; Pedreschi, D.; Giannotti, F. Stable and actionable explanations of black-box models through factual and counterfactual rules. *Data Min. Knowl. Discov*. **2024**, 38, 2825–2862.
44. Konstantinov, A.V.; Utkin, L.V. Interpretable machine learning with an ensemble of gradient boosting machines. *Knowl.-Based Syst.* **2021**, 222, 106993.
45. Ibarguren, I.; Perez, M.J.; Muguerza, J.; Arbelaitz, O.; Yera, A. PCTBagging: From inner ensembles to ensembles. A trade-off between discriminating capacity and interpretability. *Inf. Sci.* **2022**, 592, 198–217.
46. Bruggenjurgen, S.; Schaaf, N.; Kerschke, P.; Huber, M.F. Mixture of decision trees for interpretable machine learning. *arXiv* **2022**, arXiv:2211.14617.
47. Vogel, R.; Schlosser, T.; Manthey, R.; Ritter, M.; Vodel, M.; Eibl, M.; Schneider, K.A. A meta algorithm for interpretable ensemble learning: The league of experts. *Mach. Learn. Knowl. Extr*. **2024**, 6, 863–885.
48. Arwade, G.; Olafsson, S. Learning ensembles of interpretable simple structures. *Ann. Oper. Res.* **2025**.
49. Blanchart, P. An exact counterfactual-example-based approach to tree-ensemble models interpretability. *arXiv* **2021**, arXiv:2105.14820.
50. Alshkeili, H.M.H.A.; Almheiri, S.J.; Khan, M.A. Privacy-preserving interpretability: An explainable federated learning model for predictive maintenance in sustainable manufacturing and Industry 4.0. *AI* **2025**, 6, 117.
51. Haque, M.E.; Jahidul Islam, S.M.; Mia, S.; Sharmin, R.; Ashikuzzaman; Morshed, M.S.; Huque, M.T. StackLiverNet: A novel stacked ensemble model for accurate and interpretable liver disease detection. In *Proceedings of the 16th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Indore, India, 6–11 July 2025.

 doi:10.20944/preprints202604.2101.v1

52. Moreno-Sanchez, P.A. Development of an explainable prediction model of heart failure survival by using ensemble trees. In *Proceedings of the 2020 IEEE International Conference on Big Data*, Atlanta, GA, USA (Virtual), 10–13 December 2020; pp. 4902–4910.
53. Zheng, H.-L.; An, S.-Y.; Qiao, B.-J.; Guan, P.; Huang, D.-S.; Wu, W. A data-driven interpretable ensemble framework based on tree models for forecasting COVID-19. *Environ. Sci. Pollut. Res.* **2023**, 30, 66085–66103.
54. Jia, J.-F.; Chen, X.-Z.; Bai, Y.-L.; Li, Y.-L.; Wang, Z.-Y. An interpretable ensemble learning method to predict the compressive strength of concrete. *Structures* **2022**, 38, 644–655.
55. Song, Z.; Cao, S.; Yang, H. An interpretable framework for modeling global solar radiation using tree-based ensemble machine learning. *Appl. Energy* **2024**, 365, 123238.
56. Li, L.; Qiao, J.; Yu, G.; Wang, L.; Li, H.-Y.; Liao, C.; Zhu, Z. Interpretable tree-based ensemble model for predicting beach water quality. *Water Res.* **2022**, 211, 118078.
57. Qiu, Y.; Zhou, J. Short-term rockburst prediction in underground project: Insights from an explainable and interpretable ensemble learning model. *Acta Geotech.* **2023**, 18, 6655–6685.
58. Saadallah, A. Online adaptive local interpretable tree ensembles for time series forecasting. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW)*, Abu Dhabi, UAE, 9–12 December 2024; pp. 229–237.

 doi:10.20944/preprints202604.2101.v1

59. Vultureanu-Albisi, A.; Badica, C. Improving students' performance by interpretable explanations using ensemble tree-based approaches. In *Proceedings of the IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara, Romania, 19–21 May 2021; pp. 215–220.
60. Arévalo-Cordovilla, F.E.; Peña, M. ;*Sci. Rep.* **2025**, 15, 223.