# INTERPRET WHEN POSSIBLE: AN ADAPTIVE TREE-BASED HYBRID FRAMEWORK FOR EXPLAINABLE CLASSIFICATION SYSTEMS

Author's: V. Priyadharshan[1], N. Vanjulavalli[2], K. Sujith[3]

**Abstract**

This paper presents Interpret When Possible: A Tree-Based Hybrid Framework for Interpretable Classification, a novel approach that balances predictive accuracy with model interpretability. Modern machine learning systems often rely on complex black-box models that achieve high performance but lack transparency. In contrast, interpretable models such as decision trees provide clear reasoning but may sacrifice accuracy in complex decision regions.To address this trade-off, the proposed framework integrates a gating decision tree that dynamically determines whether an input instance can be reliably classified using an interpretable model or should be deferred to a high-performance black-box model. When confidence and agreement conditions are satisfied, the system produces transparent, rule-based predictions. Otherwise, it leverages the black-box model to maintain classification accuracy.Experimental evaluation on synthetic data demonstrates that the hybrid framework achieves competitive overall accuracy while significantly increasing interpretability coverage. The results highlight that selective interpretability is a practical and scalable strategy for building trustworthy AI systems without compromising predictive performance.