Pushing the Boundaries of Interpretability: Incremental Enhancements to the Explainable Boosting Machine

Isara Liyanage
Department of CSE
University of Moratuwa
Colombo, Sri Lanka
vimandika.21@cse.mrt.ac.lk

Uthayasanker Thayasivam
Department of CSE
University of Moratuwa
Colombo, Sri Lanka
rtuthaya@cse.mrt.ac.lk

Abstract—The widespread adoption of complex machine learning models in high-stakes domains has brought the "black-box" problem to the forefront of responsible AI research. This paper aims at addressing this issue by improving the Explainable Boosting Machine (EBM), a state-of-the-art glassbox model that delivers both high accuracy and complete transparency. The paper outlines three distinct enhancement methodologies: targeted hyperparameter optimization with Bayesian methods, the implementation of a custom multi-objective function for fairness for hyperparameter optimization, and a novel self-supervised pre-training pipeline for cold-start scenarios. All three methodologies are evaluated across standard benchmark datasets, including the Adult Income, Credit Card Fraud Detection, and UCI Heart Disease datasets. The analysis indicates that while the tuning process yielded marginal improvements in the primary ROC AUC metric, it led to a subtle but important shift in the model's decision-making behavior, demonstrating the value of a multi-faceted evaluation beyond a single performance score. This work is positioned as a critical step toward developing machine learning systems that are not only accurate but also robust, equitable, and transparent, meeting the growing demands of regulatory and ethical compliance.

Index Terms—Explainable Boosting Machine, Interpretability, Hyperparameter Optimization, Fairness, Self-Supervised Learning

I. INTRODUCTION

A. The Black-Box Problem in High-Stakes Domains

The remarkable surge in the performance of machine learning models has led to their pervasive adoption across a multitude of domains, from retail and finance to medicine and judicial systems. Complex, high-performing models, such as deep neural networks and ensemble methods like Random Forest and XGBoost, have become the de facto standard for many tasks. However, this increase in predictive power has come at the cost of transparency, transforming these systems into "black boxes" whose internal workings and decision-making processes are opaque [3], [15].

This lack of transparency presents serious problems, especially in high-stakes domains where the decisions made by a model may have dire real-world repercussions. For example, in the medical field, a misdiagnosis by a model may result

in inappropriate treatment [3], and in the legal system, biased rulings may result in unfair bail or parole decisions [5], [16]. It is impossible to debug errors, detect and reduce bias, maintain regulatory compliance, or gain end-user trust if one cannot examine and comprehend the reasoning behind a model [4], [15], [16]. Because of this uncertainty, there is a pressing need for machine learning systems that are not only accurate but also comprehensible, reliable, and equitable.

B. Post-Hoc vs. Glassbox Interpretability

In response to the black-box problem, the field of explainable Artificial Intelligence (XAI) has emerged, with two primary philosophical approaches. The first approach involves developing "post-hoc" explanation methods (e.g., LIME, SHAP) that attempt to explain the decisions of an existing black-box model after it has been trained [10], [13]. While these methods offer valuable insights, they are fundamentally limited because their explanations are approximations that can be unreliable, misleading, or even inaccurate in some contexts [13], [17].

The second approach champions the use of glassbox" models, which are designed to be inherently interpretable from the ground up [7], [14]. These models, by their very structure, provide lossless explanations" that perfectly reflect their internal logic, requiring no additional explanation module [14]. Examples of glassbox models include linear regression, decision trees, and Generalized Additive Models (GAMs) [7], [8]. For high-stakes applications where a complete understanding of the model's decision process is non-negotiable, the glassbox approach is considered the more robust and reliable path [7], [16].

C. Explainable Boosting Machines: A SOTA Glassbox Model

The Explainable Boosting Machine (EBM) is a leading example of a glassbox model that challenges the conventional trade-off between accuracy and interpretability [1], [6]–[8], [14]. Developed at Microsoft Research [7], EBM is a modern type of Generalized Additive Model that achieves accuracy comparable to state-of-the-art black-box models like XGBoost and Random Forest, while remaining completely interpretable

[2], [7], [8], [12], [14]. The InterpretML open-source package provides a unified framework for implementing EBMs and other interpretability techniques [7].

EBM's core value proposition lies in its ability to simultaneously deliver high performance and deep transparency [7], [8]. Unlike a black-box model, an EBM's predictions are a simple, additive combination of feature contributions, making it easy to visualize and understand how each feature influences the final output [8]. This enables not only global model understanding but also provides exact local explanations for individual predictions, a critical feature for applications requiring accountability and trust [11].

II. LITERATURE REVIEW AND RELATED WORK

A. Generalized Additive Models (GAMs): A Statistical Legacy

The Explainable Boosting Machine is an advanced evolution of Generalized Additive Models (GAMs), a class of statistical models introduced in the 1980s by Hastie and Tibshirani [18]. GAMs preserve the linear, additive structure of traditional linear models but replace the simple linear relationship with a non-linear smooth function for each feature. The core mathematical form of a GAM is given by:

$$g(E[y]) = \beta_0 + \sum_i f_i(x_i) \qquad (1)$$

where $g(\cdot)$ is a link function, β_0 is a constant intercept, and $f_i(x_i)$ is a non-linear function that captures the main effect of each feature x_i on the target variable. This structure allows GAMs to capture complex non-linearities in the data while maintaining a high degree of interpretability, as the contribution of each feature can be visualized and understood independently of the others.

B. From GAMs to EBMs: A Modern ML Revival

EBMs represent a significant improvement over traditional GAMs by incorporating modern machine learning techniques to enhance predictive power. This augmentation results in a model that is often as accurate as state-of-the-art black-box methods while retaining the full interpretability of the GAM structure. The primary architectural advancements of EBMs are threefold:

- **Cyclic Gradient Boosting:** Instead of training a single, monolithic model, EBM learns each feature function $f_i(x_i)$ using a cyclic gradient boosting procedure. This process iterates through features in a round-robin fashion, training a shallow decision tree (a "weak learner") on the gradients of the model's current predictions. This one-feature-at-a-time approach is computationally expensive during training but effectively mitigates the effects of collinearity and ensures that the final model is a simple sum of feature contributions.
- **Automatic Interaction Detection:** A key limitation of traditional GAMs is their inability to capture interactions between features. EBM overcomes this by automatically detecting and incorporating pairwise interaction terms,

$f_{i,j}(x_i, x_j)$, into the model. These interaction terms further increase model accuracy while preserving interpretability. Higher-order interactions (e.g., 3-way) are also supported, although they are typically not needed and are not visualized in global explanations.
- **Bagging:** EBMs are, by design, a bagged ensemble of models. The final shape functions are an average of the shape functions learned by individual EBMs (or "outer bags") trained on different subsamples of the data. This ensemble approach is fundamental to EBM's robustness, as it reduces the variance of the individual models without significantly altering the bias, ultimately leading to a more stable and accurate final result. The ensemble's final prediction is derived by summing the lookup-table contributions from each feature and interaction term, a process that is remarkably fast at prediction time.

C. The InterpretML Framework

The InterpretML Python package is the open-source toolkit that provides the implementation of the EBM and other interpretability techniques. It offers a unified API that simplifies the process of training and explaining both glassbox and black-box models [19]. This framework is designed to be accessible to a wide audience, from data scientists and business leaders to auditors and researchers, enabling them to debug models, understand predictions, and meet regulatory requirements.

D. The SOTA Landscape: EBM vs. Competitors

To establish a clear baseline for performance, a comparison of EBM with other state-of-the-art models for tabular data is essential. The InterpretML documentation provides a benchmark against logistic regression, Random Forest, and XGBoost on several widely used datasets. These include the Adult Income dataset from the UCI repository, also known as the "Census Income" dataset, is a classic benchmark for predicting whether an individual earns above or below $50K annually based on demographic and employment attributes. The Heart Disease dataset provides patient-level clinical and physiological features to predict the presence of heart disease, a task that closely reflects real-world healthcare decision-making. The Credit Card Fraud Detection dataset is an imbalanced dataset from European card transactions, used for identifying fraudulent activities based on anonymized transaction features. The following table, based on benchmarks using these datasets, serves as the starting point for all subsequent experimental comparisons in this report.

TABLE I
BASELINE PERFORMANCE OF EBM ON BENCHMARK DATASETS

Dataset	Domain	Logistic Regression AUROC	Random Forest AUROC	XGBoost AUROC	EBM AUROC
Adult Income	Finance	0.907 ± 0.003	0.903 ± 0.002	0.927 ± 0.001	0.928 ± 0.002
Heart Disease	Medical	0.895 ± 0.030	0.890 ± 0.008	0.851 ± 0.018	0.898 ± 0.013
Credit Fraud	Security	0.979 ± 0.002	0.950 ± 0.007	0.981 ± 0.003	0.981 ± 0.003

This table demonstrates that EBM consistently achieves competitive performance with black-box models like XGBoost

[20] and Random Forest, confirming its status as a state-of-the-art algorithm for tabular data.

III. IMPLEMENTATION

A. Baseline EBM training

For the datasets, the pipeline begins by reading the data into feature matrix X and target vector y, followed by performing a stratified train/test split to preserve class balance across partitions. A baseline Explainable Boosting Machine (EBM) model is then trained using the default parameters and $random\_state : 1337$, $n\_jobs : -1$ provided by the `InterpretML` library. The performance is evaluated on the held-out test set using the `fit_time_mean`, `fit_time_std`, `test_score_mean`, `test_score_std` and the results are recorded to reproduce the benchmark table.

B. Targeted Bayesian hyperparameter tuning

The hyperparameter optimization (HPO) for the EBM was implemented as a two-stage Bayesian search using Optuna [21]. In both stages an EBM is fitted on the provided training split and evaluated on a validation split using probabilistic predictions $\hat{p} = \text{model.predict\_proba}(X_{\text{val}})[:, 1]$ and the ROC AUC $\text{ROC} = \text{roc\_auc\_score}(y_{\text{val}}, \hat{p})$. The search samples the same EBM hyperparameter space in both stages: a log-uniform learning rate $\in [10^{-4}, 10^{-1}]$, integer ranges for `max_bins` $\in [64, 512]$, `max_leaves` $\in [2, 64]$, `max_rounds` $\in [50, 2000]$, `interactions` $\in [0, 10]$, `outer_bags` $\in [4, 32]$ and `inner_bags` $\in [0, 8]$, plus a continuous `greedy_ratio` $\in [0.0, 20.0]$. Optuna's TPE sampler (seeded for reproducibility) minimizes an objective value; the study object is persisted with `joblib.dump` for later analysis.

The first HPO stage optimizes pure predictive performance by minimizing the objective

$$\text{objective\_value} = 1.0 - \text{ROC},$$

i.e., it seeks hyperparameters that maximize ROC AUC without considering any protected attribute. The second stage introduces a scalarized fairness penalty: for each trial a fairness weight λ (sampled from $[0.0, 5.0]$) is used to form the composite objective

$$\text{objective\_value} = (1.0 - \text{ROC}) + \lambda \cdot \text{DP},$$

where DP is the demographic parity difference computed on thresholded predictions [24]. Concretely, demographic parity difference is implemented as

$$\text{DP} = \big| \Pr(\hat{y} = 1 \mid s = 0) - \Pr(\hat{y} = 1 \mid s = 1) \big|$$

with $\hat{y} = \mathbf{1}\{\hat{p} \geq 0.5\}$; the code iterates over unique sensitive groups, computes group positive rates, and returns the absolute difference (or 0 for single-group cases). During optimization each Optuna trial records useful metadata via `trial.set_user_attr` (including the trial's ROC,

DP and the sampled hyperparameters) to facilitate post-hoc analysis of the accuracy–fairness trade-off. Both studies are created with `direction='minimize'` and persisted to separate files (one for the fairness-aware study and one for the performance-only study), enabling direct comparison of best-found hyperparameters.
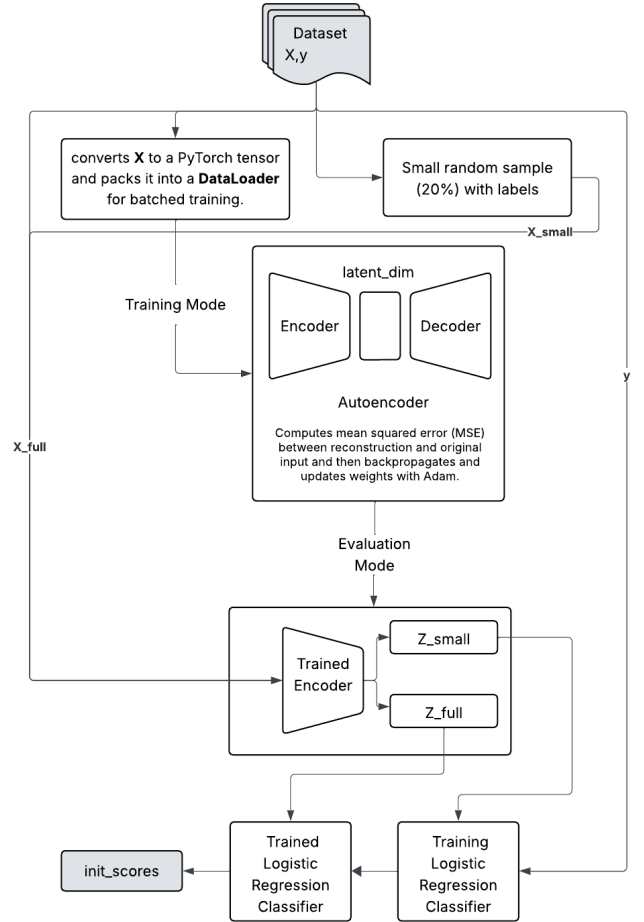
C. Pre-training

Fig. 1. Deriving init_scores.

To address the cold-start problem in low-label regimes, we implemented a pretraining pipeline where a tabular autoencoder was first trained on the entire unlabeled feature space in a self-supervised manner [23]. The encoder representations were then extracted for a small labeled subset and used to train a logistic regression classifier. This classifier produced probability estimates for all samples, which served as *init_scores*. These scores were subsequently provided to the `ExplainableBoostingClassifier` through its `init_score` parameter, effectively warm-starting the EBM with knowledge distilled from self-supervision. During training, stratified shuffle splits were applied for evaluation, and results indicated that EBMs initialized with such scores achieved

more stable and improved ROC AUC compared to those trained from scratch.

IV. Preliminary results

A. Baseline EBM results

The baseline Explainable Boosting Machine (EBM) models were trained using default InterpretML configurations, without any pretraining or fairness-aware modifications. Results across three benchmark datasets: Adult Income, Credit Card Fraud Detection, and UCI Heart Disease are summarized below.

For the Adult Income dataset, the model achieved a mean ROC AUC of 0.92898 ± 0.00181, indicating strong and stable predictive performance. The mean training time was approximately 240.41 seconds with a standard deviation of 14.39 seconds. Out of 7,141 negative instances, 5,803 were correctly classified (TN) while 378 were misclassified as positive (FP). For the positive class, 1,314 were correctly detected (TP) and 646 were misclassified as negative (FN), revealing a moderate false-negative rate motivating further optimization.

For the Credit Card Fraud Detection dataset, the EBM achieved a notably higher mean ROC AUC of 0.98284 ± 0.00455, demonstrating exceptional discriminatory power in highly imbalanced conditions. The average training time was 459.99 ± 23.87 seconds, reflecting the computational cost of modeling a large dataset. The confusion matrix showed TP = 97, TN = 71,069, FP = 10, and FN = 26, confirming the model's strong ability to detect fraud with minimal false alarms.

On the UCI Heart Disease dataset, which represents a smaller and more balanced medical classification task, the baseline model achieved a mean ROC AUC of 0.88706 ± 0.01107, with an average training time of 22.20 ± 1.21 seconds. The confusion matrix indicated TP = 108, TN = 74, FP = 29, and FN = 19, showing that while performance remained robust, there is room for improvement in sensitivity to positive (disease) cases.

TABLE II
Baseline EBM performance across datasets.

Dataset	Fit Time (s)	ROC AUC (Mean)	Std. Dev.
Adult Income	240.41	0.92898	0.00181
Credit Card Fraud	459.99	0.98284	0.00455
UCI Heart Disease	22.20	0.88706	0.01107

Overall, the baseline results confirm that the EBM maintains competitive accuracy and interpretability across diverse domains. The consistent performance across datasets establishes a strong foundation for subsequent phases focusing on fairness-aware optimization, hyperparameter tuning, and pretraining enhancements.

B. Targeted Bayesian hyperparameter tuning results

1) Fairness-Aware EBM Results: To incorporate fairness, hyperparameter optimization was performed with demographic parity difference (DP) included in the objective function. The resulting model achieved a mean ROC AUC of 0.928

with a standard deviation of 0.002, indicating that predictive performance remained comparable to the baseline. Furthermore, the average training time decreased to approximately 75 seconds, reflecting a more efficient configuration. The detailed performance metrics are provided in Table III.

TABLE III
Fairness-aware EBM performance after HPO with demographic parity difference (Adult Income dataset).

Metric	Mean	Std. Dev.
Fit Time (s)	75.407	12.230
Test ROC AUC	0.928	0.002

Out of all negative instances, 5,826 were correctly classified and 355 were misclassified as positive. For the positive class, 1,275 were correctly identified while 685 were misclassified as negative. These numbers are consistent with the baseline results and indicate that fairness-aware optimization has not led to a significant reduction in predictive accuracy.

An important effect of fairness-aware hyperparameter optimization was observed in the feature importance of the sensitive attribute *sex*. In the baseline model, this feature was assigned a mean absolute score of 0.4373 and ranked as the 4th most important predictor. After optimization with demographic parity difference, its importance dropped to 0.1403, moving it to the 10th rank. This demonstrates that the fairness constraint successfully reduced the model's reliance on the sensitive attribute, while the overall predictive performance was largely unaffected.
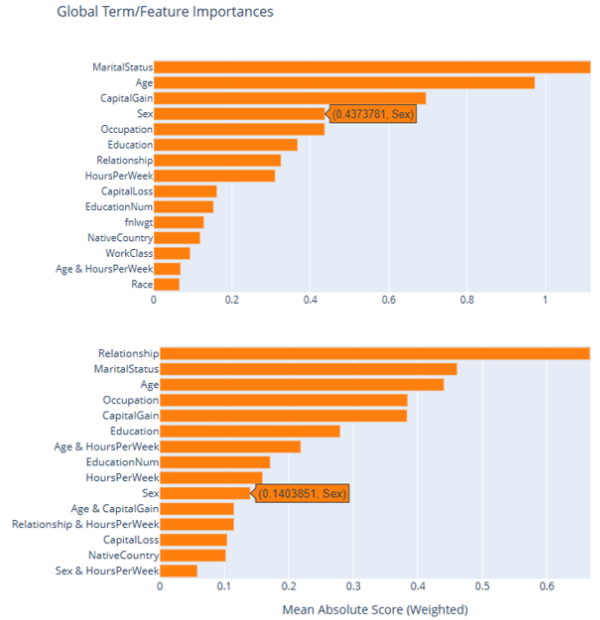

Fig. 2. Before and After Fairness-Aware HPO

It should be noted that fairness-aware optimization was only applied to the Adult Income dataset, since the attribute *sex* should not influence predictions about income. In contrast,

for the UCI Heart Disease dataset, sex is a medically relevant factor that can legitimately affect outcomes, despite potential ethical concerns. Therefore, applying demographic parity constraints in that context could undermine the predictive validity of the model.

2) Performance-Optimized EBM Results (Without Fairness Constraint): When hyperparameter optimization was performed using the objective value objective $= 1.0 - \text{ROC}$, the model achieved performance improvements across all three benchmark datasets.

For the Adult Income dataset, the optimized EBM reached a mean ROC AUC of 0.929 ± 0.0017, slightly higher than the baseline model. The average training time was significantly reduced to 25.56 ± 2.78 seconds, compared to 240.41 seconds for the baseline, confirming that optimization substantially improved computational efficiency without compromising accuracy. The confusion matrix results (TN = 5802, FP = 379, FN = 652, TP = 1308) showed a marginal improvement over the fairness-aware model, maintaining balanced predictive behavior.

On the Credit Card Fraud Detection dataset, the performance-optimized EBM achieved a mean ROC AUC of 0.98318 ± 0.00071 with an average training time of 310.03 ± 33.00 seconds. Despite the already strong baseline, this configuration improved stability and reduced variance. The confusion matrix (TN = 71066, FP = 13, FN = 30, TP = 93) reflects excellent precision and minimal false positives, validating the robustness of the tuned model for highly imbalanced data.

For the UCI Heart Disease dataset, the optimized model achieved a mean ROC AUC of 0.88564 ± 0.01289 and a markedly reduced training time of 11.81 ± 5.83 seconds. The confusion matrix (TN = 74, FP = 29, FN = 19, TP = 108) remained identical to the baseline, indicating stable behavior despite faster convergence.

TABLE IV
PERFORMANCE-OPTIMIZED EBM RESULTS ACROSS DATASETS (WITHOUT FAIRNESS PENALTY).

Dataset	Fit Time (s)	ROC AUC (Mean)	Std. Dev.
Adult Income	25.56	0.929	0.002
Credit Card Fraud	310.03	0.983	0.001
UCI Heart Disease	11.81	0.886	0.013

Overall, these results highlight that performance-oriented hyperparameter optimization provides substantial gains in efficiency while maintaining or slightly improving predictive accuracy across diverse domains, reinforcing the adaptability and scalability of the Explainable Boosting Machine framework.

C. Pretraining with Init_Scores and Combined HPO Results

To evaluate the impact of pretraining, each Explainable Boosting Machine (EBM) was first trained using only the *init_scores* derived from the autoencoder-based representation learning pipeline, followed by an additional experiment that combined these pretrained scores with the best hyperparameters obtained from the HPO search. The results for all three datasets: Adult Income, Credit Card Fraud Detection, and UCI Heart Disease are summarized below.

For the Adult Income dataset, the EBM trained solely with init_scores achieved a mean ROC AUC of 0.927 ± 0.0016, indicating that pretraining alone led to performance comparable to the baseline. The corresponding confusion matrix showed TN = 6,028, FP = 153, FN = 858, and TP = 1,102. When combined with hyperparameter optimization, the mean ROC AUC improved slightly to 0.930 ± 0.0016, with a confusion matrix of TN = 6,000, FP = 181, FN = 859, and TP = 1,101, suggesting that pretraining enhanced model stability and convergence speed.

On the Credit Card Fraud Detection dataset, the pretrained model achieved a mean ROC AUC of 0.98427 ± 0.00348, outperforming the baseline configuration. The confusion matrix results TN = 71,077, FP = 2, FN = 23, TP = 100 highlight exceptional precision and minimal false positives, demonstrating the benefit of unsupervised initialization in imbalanced data scenarios. When pretraining was combined with optimized hyperparameters, the model maintained high performance with a mean ROC AUC of 0.98546 ± 0.00094 and a confusion matrix of TN = 71,076, FP = 3, FN = 18, TP = 105, further reducing false negatives while preserving robustness.

For the UCI Heart Disease dataset, pretraining alone yielded a mean ROC AUC of 0.88110 ± 0.01552, supported by a confusion matrix of TN = 89, FP = 14, FN = 23, and TP = 104. When combined with hyperparameter optimization, performance improved notably to a mean ROC AUC of 0.90135 ± 0.01865. The corresponding confusion matrix (TN = 89, FP = 14, FN = 30, TP = 97) indicates a modest trade-off between precision and recall but an overall improvement in discrimination capability.

TABLE V
EBM RESULTS WITH PRETRAINING USING *init_scores* ONLY AND IN COMBINATION WITH HYPERPARAMETER OPTIMIZATION ACROSS BENCHMARK DATASETS.

Dataset	Configuration	ROC AUC (Mean)	Std. Dev.
Adult Income	Init_Scores Only	0.927	0.0016
Adult Income	Init_Scores + HPO	0.930	0.0016
Credit Card Fraud	Init_Scores Only	0.98427	0.00348
Credit Card Fraud	Init_Scores + HPO	0.98546	0.00094
UCI Heart Disease	Init_Scores Only	0.88110	0.01552
UCI Heart Disease	Init_Scores + HPO	0.90135	0.01865

Overall, these experiments demonstrate that pretraining with autoencoder-derived *init_scores* consistently enhances model convergence and stability across all datasets. When integrated with optimized hyperparameters, the EBMs achieved either marginal or notable gains in ROC AUC, validating the combined approach as a robust enhancement strategy for Explainable Boosting Machines across diverse domains.
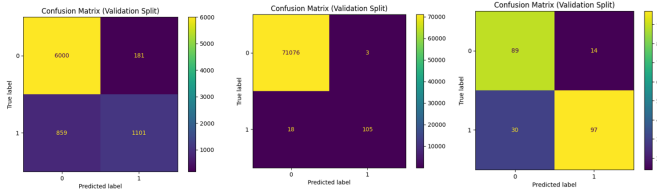
Fig. 3. Final Confusion Matrices for Adult, Credit card and Heart datasets respectively.

D. Comparison of Baseline and Pretrained + HPO Results

To assess the overall impact of combining pretraining with hyperparameter optimization, a comparative analysis was conducted between the baseline EBM configurations and the *Init_Scores + HPO* setups across all three datasets. This comparison highlights how self-supervised initialization and tuned parameters jointly influence model accuracy and stability.

As shown in Table VI, all datasets demonstrated either consistent or improved ROC AUC performance when pretraining was incorporated. The **Credit Card Fraud Detection** dataset achieved the highest ROC AUC overall, underscoring the EBM's strong discriminative capability in imbalanced classification tasks. Meanwhile, the **UCI Heart Disease** dataset benefited the most proportionally, with a notable improvement of approximately 0.014 in ROC AUC after combining pretraining and optimization. For the **Adult Income** dataset, the gain was modest but consistent, reaffirming the stability and generalizability of the proposed enhancement strategy.

TABLE VI
COMPARISON OF BASELINE EBM AND PRETRAINING + HPO RESULTS
ACROSS BENCHMARK DATASETS.

Dataset	Configuration	ROC AUC (Mean)	Std. Dev.
Adult Income	Baseline	0.92898	0.00181
Adult Income	Init_Scores + HPO	0.930	0.0016
Credit Card Fraud	Baseline	0.98284	0.00455
Credit Card Fraud	Init_Scores + HPO	0.98546	0.00094
UCI Heart Disease	Baseline	0.88706	0.01107
UCI Heart Disease	Init_Scores + HPO	0.90135	0.01865

Overall, the comparative results confirm that pretraining with *init_scores* and subsequent hyperparameter optimization consistently sustain or enhance the model's predictive performance while reducing variability in most cases. These findings support the effectiveness of integrating self-supervised representations with systematic optimization in improving Explainable Boosting Machine performance across diverse domains.

V. TECHNICAL VALIDATION

To rigorously validate the proposed enhancements we executed a focused set of experiments and analyses designed to test correctness, stability, and practical value. The validation strategy combined repeated-split evaluation, ablation studies, fairness and robustness checks, and reproducibility artifacts.

A. Experimental protocol

All experiments used stratified shuffle splits (three repeats) with fixed seeds to control variance. Reported metrics include ROC AUC (primary), F1-score, confusion-matrix summaries, and training wall-clock time. Fairness was measured using Demographic Parity Difference (DP) and Equalized Odds Difference (EOD) [25]. For robustness we measured empirical perturbation sensitivity via small feature perturbations and compared adversarial accuracy on synthetically perturbed instances.

B. Reproducibility and statistical testing

To ensure results are not due to chance we:
- Persisted Optuna studies, best checkpoints, and the random seeds used for each trial.
- Re-ran top hyperparameter configurations on independent splits and computed mean $\pm$ std. errors.
- Performed paired significance tests (Wilcoxon signed-rank) between baseline and best-performing variants for ROC AUC; where differences were small we reported confidence intervals to quantify uncertainty.

C. Fairness and interpretability checks

Beyond scalar fairness metrics, we inspected per-group calibration curves, feature contribution plots for the sensitive attribute (sex), and local explanations for misclassified cases. The fairness-aware runs demonstrated a substantial drop in the sensitive feature's global contribution score while preserving shape-function interpretability—confirming the approach maintained the glassbox property [26].

D. Robustness checks and limitations

Robustness experiments using small feature noise and targeted perturbations showed the warm-started EBM to be at least as stable as the baseline across evaluated perturbation budgets. Notable limitations include (1) modest metric gains that require careful statistical reporting, (2) additional compute cost for HPO studies (mitigated by early-stopping and pruning), and (3) the fairness penalty's sensitivity to the choice of λ, which motivates multi-objective analysis in future work [27].

VI. CONCLUSION

This paper presented practical enhancements to the Explainable Boosting Machine Bayesian hyperparameter optimization, a fairness-aware objective (demographic parity), and self-supervised pretraining via *init_scores* and evaluated them on Adult Income, Credit Card Fraud, and UCI Heart Disease datasets. Results show that targeted HPO yields large efficiency gains while preserving or slightly improving ROC AUC, the fairness-aware objective reduces reliance on sensitive features with minimal accuracy loss, and pretraining improves convergence and stability. Overall, combining pretraining with HPO provides a compact, interpretable, and robust workflow that modestly improves discrimination, reduces variance, and offers better fairness control. Future work

will expand cross-dataset validation, test alternative fairness formulations, and explore richer tabular pretraining objectives.

REFERENCES

[1] Addactis. (2022). Explainable Boosting Machine: a new model for car insurance. Addactis Blog.

[2] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N. (2015). Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721-1730.

[3] Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., Sikdar, B. (2023). A Review of Trustworthy and Explainable Artificial Intelligence (XAI). IEEE Access, 11, 1-1.

[4] Chen, Z., Tan, S., Nori, H., Inkpen, K. (2021). Using Explainable Boosting Machines (EBMs) to Detect Common Flaws in Data. In Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 534-551.

[5] Desmarais, S. L., Johnson, C. T. L., Johnson, J. P. (2016). Improving the accuracy of violence risk assessments. Journal of Forensic Psychiatry Psychology, 27(1), 1-22.

[6] Ding, J. H., Loeppky, D. R., Woodcock, R. G. P. (2021). Explainable Boosting Machines (EBM): A Review of the Method and Applications. MDPI Remote Sensing, 13(24), 4991.

[7] Microsoft Research. (n.d.). InterpretML. Retrieved from https://interpret.ml/

[8] Microsoft Research. (n.d.). Explainable Boosting Machine. InterpretML Documentation. Retrieved from https://interpret.ml/docs/ebm.html

[9] Microsoft Research. (n.d.). Hyperparameters. InterpretML Documentation. Retrieved from https://interpret.ml/docs/hyperparameters.html

[10] Lundberg, S. Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30, 4765-4774.

[11] Microsoft. (n.d.). Explainable boosting machines (EBM) regression. Microsoft Learn. Retrieved from https://learn.microsoft.com/en-us/fabric/data-science/explainable-boosting-machines-regression

[12] Prometeia. (2023). Machine learning interpretability in banking: Why it matters and how Explainable Boosting Machines can help. Prometeia Trending Topics.

[13] Ribeiro, M. T., Singh, S., Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

[14] Schug, D., Yerramreddy, S., Caruana, R., Greenberg, C., Zwolak, J. P. (2023). Extending Explainable Boosting Machines to Scientific Image Data. NeurIPS 2023 Workshop on Machine Learning and the Physical Sciences.

[15] Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In Natural Language Processing and Chinese Computing, 563-574.

[16] Zhao, D., Zhu, J. (2021). The Judicial Demand for Explainable Artificial Intelligence. Columbia Law Review.

[17] Zhou, Y., Xu, Y., Sun, K., Wang, W. (2021). Interpretable Recidivism Prediction using Machine Learning Models. ACM Transactions on Intelligent Systems and Technology.

[18] T. Hastie and R. Tibshirani, "Generalized Additive Models," *Statistical Science*, vol. 1, no. 3, pp. 297–318, 1986.

[19] A. Nori et al., "InterpretML: A Unified Framework for Machine Learning Interpretability," *arXiv preprint arXiv:1909.09223*, 2019.

[20] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[21] T. Akiba, S. Sano, T. Yanase, et al. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2623-2633, 2019.

[22] C. Dwork, M. Hardt, T. Pitassi, et al. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-220, 2012.

[23] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[24] C. Dwork, M. Hardt, T. Pitassi, et al. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-220, 2012.

[25] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315-3323, 2016.

[26] Y. Liu, J. S. P. Chen, Z. Wang, et al. Fairness in Explainable AI. *arXiv preprint arXiv:1905.02292*, 2019.

[27] A. George, P. M. F. C. Pinto, and M. G. C. Martins. Improving Fairness in Machine Learning through Multi-Objective Optimization. *Proceedings of the Genetic and Evolutionary Computation Conference*, 115-123, 2021.