# Thesis Proposal: Sparse & Higher-Order Explainable Boosting Machines: 3-Way FAST Ranking and Main-Effect Selection

Ata Berk Çakır

September 2025

## 1 Introduction and Motivation

Explainable Boosting Machines (EBMs) are glass-box models that learn an additive decomposition of the prediction function,

$$g(x) = \beta_0 + \sum_j f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k),$$

where $g$ is a link function (identity for regression, logistic for classification), $f_j$ are one-dimensional shape functions, and $f_{jk}$ optional pairwise interaction surfaces. Shapes are trained by gradient boosting of shallow decision trees and displayed as intelligible plots, contrasting with black-box ensembles whose explanations are post-hoc [1, 4, 3].

Two challenges motivate this work. **Interaction depth gap:** most toolchains rank and train only pairwise interactions; genuine 3-way effects (e.g., gene×drug×dose, feature×time×context) remain invisible or leak into distorted pairwise proxies [2]. **Main-effect overload:** default EBMs keep shapes for all features, which clutters explanations and can induce mild overfitting on wide data.

These gaps matter in regulated and high-stakes domains where human factors (term count, visual clarity, monotonicity) are as important as accuracy. We seek to **balance predictive performance with cognitive interpretability and model parsimony**.

**Synthetic-first rationale.** Before real-data benchmarking, we use synthetic datasets with known data-generating processes (DGPs) to test faithful recovery of ground-truth main effects and higher-order interactions; ablations across signal-to-noise ratio (SNR), dimensionality, sample size, and distribution shift provide controlled stress tests uncommon in interpretability research.

## 2 Previous and Related Work

**Additive models & EBMs.** GAMs provide additive decompositions amenable to visualization and constraints [1]. GA$^2$M extends GAMs with pairwise interactions to capture non-additivities while remaining intelligible [2]. EBMs implement GA$^2$M using bagged, shallow boosted trees to learn discretized shapes with robust calibration [3].

**Interaction discovery & ranking.** Pairwise candidate ranking with FAST is a practical heuristic for screening interactions before fitting [2]. We generalize this idea to 3-way terms while accounting for lower-order effects; surrogate scoring can leverage gradients of ALE [8].

**Sparsity in additive models.** Lasso promotes term sparsity [5]; group lasso encourages per-feature selection [6]; fused lasso/total-variation penalties induce piecewise-constant or smooth shapes appropriate for discretized functions [7].

**Interpretability evaluation.** Beyond accuracy, interpretability depends on shape stability, number of terms, monotonicity adherence, and user factors. Calibration and boosting foundations follow [4, 3].

# 3 Goals and Objective

**Hypothesis.** An EBM extended with computationally tractable 3-way FAST-style interaction ranking and sparse main-effect selection can (i) match or exceed vanilla EBM accuracy while (ii) reducing the number of displayed terms and (iii) faithfully recovering true effects on synthetic data—and (iv) maintain competitive performance on real benchmarks.

### Objectives

1. Re-implement a controllable EBM training loop with minimal dependencies.
2. Design & implement 3-way interaction ranking (generalizing FAST) with hierarchical screening and pruning for tractability.
3. Introduce sparsity over main effects via a single control knob ($\lambda$ or target #active shapes), supporting L1 / group / TV-style penalties.
4. **Synthetic-first evaluation:** verify recovery of known DGP shapes/interactions under varied SNR, dimensionality, missingness, and drift.
5. **Real-data benchmarking:** compare against reference EBMs, spline-GAM, and GBDT+SHAP on accuracy, calibration, and interpretability.
6. Deliver a reproducible codebase, DGP generators, experiment scripts, and documentation.

# 4 Material and Methods

## 4.1 Data & Tasks (Synthetic first, then Real)

**Synthetic phase (primary).** We generate classification (logistic link) and regression (identity link) tasks with:

- Nonlinear main shapes (monotone, piecewise, sigmoid, spline-like), and planted pairwise and 3-way interactions (XOR-like, multiplicative, thresholded, region-specific).
- Mixed numeric/categorical features; controlled SNR; heteroskedastic noise; missingness mechanisms (MCAR/MAR); optional covariate drift.
- Dataset regimes: $p \in \{20, 50, 100\}$, $n \in \{10{,}000, 50{,}000\}$. Seeds and splits are published.

**Real-data phase (benchmarking).** Select 2–4 public tabular datasets (classification and regression) diverse in feature types and suspected interactions. Preprocessing: missingness handling, categorical encoding, stratified splits, fixed seeds.

## 4.2 Baselines

a) **Vanilla EBM** (reference implementation) with default pairwise FAST and no sparsity [3].

b) **Spline-GAM** to chart additive-only performance [1].

c) **GBDT + SHAP** as a strong non-glass-box accuracy baseline (boosting foundations [4]).

d) **Linear/Logistic** baselines for calibration sanity checks.

### 4.3 From-Scratch EBM (minimal dependencies)

- **Boosting loop:** Train shallow trees (depth 1–2) on residuals to update discretized shapes $f_j$, $f_{jk}$, $f_{jkl}$ with learning-rate schedule, bagging/subsampling, and early stopping.
- **Shape parameterization:** Quantile-binned stepwise functions; optional isotonic/spline smoothing; optional monotonic constraints via constrained bin updates.
- **Calibration:** Optional post-hoc calibration for classification.
- **Engineering:** Vectorized NumPy kernels; minimal external dependencies.

## 5 Approach

### 5.1 Pairwise & 3-Way Interactions

**Screening and candidate sets.** Hierarchical screening mitigates combinatorics:

1. Rank **main effects** by marginal residual reduction; keep top $M$.
2. Form **pairwise candidates** among top-$M$ features; score via FAST-style marginal improvement on held-out residuals; keep top $K_2$ [2].
3. Propose **3-way candidates** by augmenting top pairs with neighbors from top-$M$ and limited random exploration; keep top $K_3$.

**3-Way FAST generalization (scoring).** For a triple $(j, k, \ell)$, define

$$S_{jkl} \;=\; \Delta\mathcal{L}\Big(\hat{f}_{jkl}\,\Big|\,r\Big) \;-\; \sum_{\text{subsets}} \Delta\mathcal{L}\Big(\hat{f}_{\text{subset}}\,\Big|\,r\Big)\,,$$

where $r$ are current residuals and $\Delta\mathcal{L}$ is validation loss reduction from fitting a coarsely binned ternary surface, subtracting lower-order gains to estimate a *pure* 3-way contribution. We approximate $\hat{f}_{jkl}$ using tiny depth-1/2 stumps over a tri-grid or ALE-gradient surrogates [8].

**Pruning and selection.** Stop if $S_{jkl} < \varepsilon$; require a permutation-null percentile (e.g., 95th); cap $K_2$, $K_3$, and bins; drop candidates with low support mass. Selected 3-way terms are fit inside boosting and regularized similarly to pairwise terms.

### 5.2 Sparse Main-Effect Selection

Two routes behind a **single control knob**:

- **Route A (penalized):** *Group lasso* at the per-feature level for on/off selection [6], with *fused lasso/TV* on bins for piecewise-constant shapes [7]; implement proximal updates; tune $\lambda$.
- **Route B (post-hoc):** Prune shapes by global contribution (validation gain $\times$ prevalence or average $|f_j|$) to a threshold minimizing validation loss.

Expose $\lambda$ or a target number of active shapes; report Pareto curves (sparsity vs. accuracy).

### 5.3 Evaluation Protocol

**Synthetic (ground-truth-aware).**

- **Shape recovery:** Bin-aligned MSE/MAE vs. true $f_j$; Earth-Mover's Distance between effect histograms.

- **Interaction recovery:** Precision/Recall@k of discovered pairs/3-ways vs. planted sets; permutation significance.
- **Ranking quality:** AUROC/AUPRC against the planted set.
- **Prediction & calibration:** LogLoss/Brier (classification), RMSE/MAE (regression).
- **Stability:** Correlation of shapes across resamples; Jaccard of selected term sets.
- **Efficiency:** Wall-clock time, candidate counts, memory.

**Real-data benchmarking.** Same predictive and interpretability metrics; plus term count, average bins/shape, and monotonicity violations. Baselines: vanilla EBM, spline-GAM, and GBDT+SHAP.

### 5.4 Deliverables

Clean codebase; DGP generators; experiment scripts; configs; seeds; reproducibility checklist; a user guide for interpreting shapes and interaction plots.

## 6 Time Plan

| Date | Milestone |
|---|---|
| 2025-07-30 | Kickoff (done); initial alignment and scoping. |
| 2025-09-24 | Literature map; finalize synthetic DGP specs seeds. |
| 2025-10-15 | EBM scaffolding (losses, binning, boosting); main-effects baseline + unit tests. |
| 2025-11-05 | Pairwise screening (FAST-style) and fitting; validate pairwise ranking on synthetic. |
| 2025-11-26 | 3-way ranking prototype; pruning rules and complexity caps; micro-benchmarks. |
| 2025-12-17 | Integrate top-m 3-way terms into boosting; synthetic validation on planted triples. |
| 2026-01-07 | Sparse main-effect selection (group lasso + fused lasso TV); proximal updates; Pareto curves. |
| 2026-01-28 | Integration synthetic ablations (SNR, n, p, missingness, drift); efficiency pass. |
| 2026-02-18 | Transition to real datasets; preprocessing; baseline runs (GAM, GBDT+SHAP, vanilla EBM). |
| 2026-03-11 | Full benchmarking; results synthesis; draft writing feedback. |
| 2026-04-01 | Final writing; reproducibility pack (code, configs, seeds); submission assets. |

## References


[1] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. Chapman & Hall/CRC, 1990.

[2] Y. Lou, R. Caruana, and J. Gehrke, "Accurate Intelligible Models with Pairwise Interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013, pp. 623–631.

[3] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A Unified Framework for Machine Learning Interpretability," *arXiv:1909.09223*, 2019.

[4] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[5] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[6] M. Yuan and Y. Lin, "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[7] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 1, pp. 91–108, 2005.

[8] D. W. Apley and J. Zhu, "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," *Journal of the Royal Statistical Society: Series B*, vol. 82, no. 4, pp. 1059–1086, 2020.