IJBITA

International Journal of Business Integrity and Technology Advancements (IJBITA)

Volume – 02
Issue – 01
(2026)

Explainable and Data Efficient Machine Learning for Financial Fraud Detection in Resource Constrained Institutions

Md Saiful Islam, College of Graduate and Professional Studies, Trine University, Angola, IN 46703, USA.
saifulcu105@gmail.com
Syed Azazul Haque, Masters of Business Administration, Department of Business Analytics, International American University Los Angeles, azazbos@gmail.com
Omar Faruq, College of Business, University, Westcliff University, Los Angeles, USA
o.faruq.123@westcliff.edu
Md. Rezaul Hasan, Department of Management, Richards College of Business, University of West Georgia, Carrollton, GA 30118, USA.
Ehasan1@my.westga.edu
Masrul Hasan, College of Graduate and Professional Studies, Trine University, Angola, USA.
mhasan23@my.trine.edu

Article Info:
Received: 23 Jan, 2026
Processed: 27 Feb, 2026
Published: 25 Mar, 2026

Keywords:
financial fraud detection; explainable machine learning; data efficiency; resource constraints; chronological evaluation; average precision; explanation stability.

CC BY SA

ABSTRACT

Fraud detection systems must balance scarce labels, alert workload, computing resources, and the information conveyed by explanations. This study compares logistic regression, a shallow decision tree, an additive Explainable Boosting Machine (EBM), and histogram gradient boosting on the public ULB and Worldline credit card benchmark. Exact full row deduplication retained 283,726 transactions with 473 fraud labels. Timestamp preserving chronological partitions contained 170,235 training, 56,745 validation, and 56,746 test transactions; the test period included 74 frauds. Nested samples of 10, 25, 50, and 100 percent of the historical training pool were evaluated across five seeds, producing 80 primary fits and one interaction ablation. Fixed configurations used two CPU threads without hyperparameter search. At full data, mean test average precision was 0.7957 for histogram boosting, 0.7512 for additive EBM, 0.7442 for logistic regression, and 0.6488 for the shallow tree. Histogram boosting recovered 86.49 percent of frauds when ranking the highest scoring 0.5 percent of the complete test batch. Frozen validation thresholds produced different future alert volumes. Full budget mean fitting times ranged from 0.40 to 35.03 seconds, while prediction and explanation costs followed different rankings. A low budget EBM fit returned constant zero probabilities and remained in the primary summaries; its diagnostic repeat reproduced the outcome. Explanation reconstruction and within model feature overlap were evaluated without treating them as evidence of causal meaning or investigator usefulness. Independent recalculation verified all 81 saved prediction records. The findings characterize configured learners on one short, anonymized benchmark and provide a transparent basis for institution specific replication.

STUDY SCOPE

The numerical results derive from executed fits on public transaction data. No financial institution supplied staffing limits, deployment hardware, investigation outcomes, or loss prevention evidence. The intended application motivates the questions; it does not define the empirical sampling frame.

Corresponding email: saifulcu105@gmail.com

INTRODUCTION

A fraud model can rank transactions effectively while remaining unsuitable for a particular institution. Limited labeled cases affect fitting, a small investigation team limits alert volume, and a strict decision deadline limits feature retrieval and scoring. An explanation can also be numerically faithful without communicating an understandable reason for a decision. Evaluating these properties together requires distinct measurements rather than a single accuracy figure.

This study asks how reduced historical label availability affects detection, computational cost, and explanation consistency across interpretable models and a tree ensemble. The experiment compares four configured learners on the same chronological test period. It varies training sample size within a fixed historical window, retains failed fits, and distinguishes a validation selected threshold from retrospective test batch ranking.

Established research motivates the comparison. Grinsztajn et al. (2022) report strong performance of tree based learners across a broad tabular benchmark. Nori et al. (2019) distinguish models designed for inspection from methods that explain an existing model. Rudin (2019) argues for considering interpretable models in consequential decisions. These sources justify including transparent baselines without presuming that one model family must dominate this fraud dataset.

The intended setting is an institution with limited labels or computing capacity. Institutional size is not a measured predictor of those limitations. A community bank, a specialist payments provider, and a larger organization can face different constraints, but none supplied requirements for this experiment. Resource constraints are therefore represented by declared CPU execution, fixed model configurations, labeled training fractions, and explicit alert budgets.

The contribution is empirical traceability. The study connects a documented protocol to individual fit outputs, saved test probabilities, contribution arrays, and independent numerical checks. It provides a descriptive comparison rather than a new learning algorithm, an optimized search across all architectures, or a claim of production effectiveness. Interpretable model structure, data efficient ranking, small serialized size, and useful analyst explanations remain separate propositions.

The sections that follow review related work, define the data and estimands, specify models and measurements, report the verified results, and identify the evidence required for institutional replication. The evaluation remains useful if its conclusion is that a simple baseline is preferable under one constraint and a more complex learner is preferable under another.

RELATED RESEARCH AND CONTRIBUTION BOUNDARIES

Realistic fraud modeling must account for imbalanced outcomes and delayed knowledge. Dal Pozzolo et al. (2018) address a practical credit card fraud learning strategy. Carcillo et al. (2018) describe streaming detection under imbalance, changing behavior, and feedback latency; Carcillo et al. (2021) examine the combination of supervised and unsupervised information. These studies support an operationally explicit evaluation, while their results cannot be directly ranked against a different split and configuration set.

Explainability research also separates several goals. Barredo Arrieta et al. (2020) organize concepts and audiences in explainable AI. Lundberg and Lee (2017) formulate SHAP as an additive feature attribution framework, and Lundberg et al. (2020) develop efficient explanations for tree models. An attribution can describe a predictor's behavior without revealing the original business meaning of an anonymized principal component.

Relevant 2026 work precludes broad novelty claims. Fazel et al. (2026) present an optimized EBM workflow involving preprocessing, feature selection, and Taguchi based optimization. The work is an arXiv preprint. Gamal et al. (2026) examine hybrid fraud learners, learning dynamics, computational cost, and SHAP and LIME explanations. Zuberi et al. (2026) study chronological fraud evaluation and explicitly distinguish short horizon variation from production drift on a much longer record.

The present experiment adopts a narrower design: four fixed configurations, four fractions of one historical training pool, five seeds, and one later test cohort. The models share training records at each seed and budget. The protocol measures ranking, threshold transport, computing workloads, and within model contribution overlap. This combination defines the comparison; it does not establish that earlier studies omitted every one of its elements.

Evaluation methodology supplies additional boundaries. Saito and Rehmsmeier (2015) motivate precision recall analysis under imbalance. Kaufman et al. (2012) and Kapoor and Narayanan (2023) explain why information leakage can invalidate a seemingly strong result. The current study limits fitted transformations to the training data but cannot reconstruct the upstream PCA fitting population supplied by the dataset creators.

Corresponding email: saifulcu105@gmail.com

The literature is a targeted selection checked on November 12, 2026, rather than a systematic review. Published titles, venues, and DOI records are documented in the references and supplement. Comparisons with other studies remain conceptual unless preprocessing, populations, metrics, and policy rules are genuinely aligned.

SCOPE AND INSTITUTIONAL CONSTRAINTS

Data efficiency is defined as the behavior of a configured supervised learner when fewer labeled records from the same historical training interval are available. The experiment changes which records are fitted; it does not shorten the interval or simulate the cost of acquiring labels. Unlabeled records are not used by a semi supervised component.

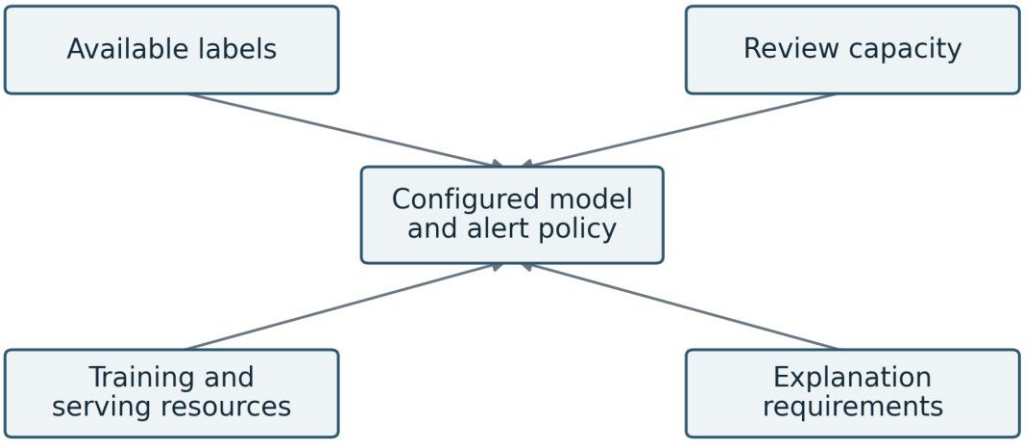


Fig 1: Four dimensions of the intended decision problem. Label availability, review capacity, computing resources, and explanation requirements constrain a configured learner and its action policy. The diagram is a design framework; these institutional constraints were not measured in participating organizations.

Detection performance concerns ranking and operational errors. Computing cost concerns defined training, prediction, and explanation workloads. Explanation consistency concerns selected contribution sets across fitted models. Human understanding concerns whether a person can use the explanation for a task. Only the first three dimensions receive measurements here; no investigator study was conducted.

The principal estimand is average precision for a specified learner and labeled training fraction on the common later test cohort. Secondary estimands include fraud recall under a fixed validation selected threshold, recall under complete batch ranking, and workload specific timing. A comparison at equal training records controls information exposure without equalizing optimization effort, parameter count, or model expressiveness.

The study deliberately avoids converting label fractions into investigator hours or model timing into service guarantees. Neither mapping is identified by the public data. Instead, the protocol supplies quantities an institution could combine with its own measured requirements in a subsequent replication. A model should be selected against an actual decision constraint rather than against the broad label of a resource constrained institution.

MATERIALS AND METHODS

A. DATASET PROVENANCE AND INTEGRITY

The source is the ULB and Worldline credit card fraud benchmark, distributed by the ULB Machine Learning Group through Kaggle and corroborated by OpenML dataset 1597, version 1 (OpenML, n.d.; ULB Machine Learning Group, n.d.). The source documentation identifies European cardholder transactions from September 2013 over approximately two days. OpenML requests citation of Dal Pozzolo et al. (2015). The dataset contains transaction fraud labels, not independently adjudicated account takeover episodes.

Corresponding email: saifulcu105@gmail.com

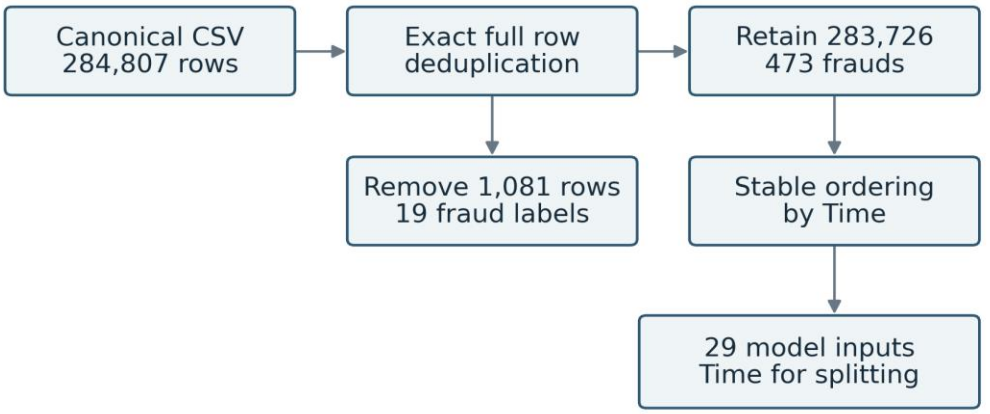


Fig 2: Data integrity pipeline and observed counts. The canonical CSV contains 284,807 transactions and 492 fraud labels. Exact full row deduplication removes 1,081 rows, including 19 fraud labeled rows, before chronological partitioning. Duplicate rows are an equality property of the supplied fields, not a verified event identity.

The analyzed CSV is identified by the SHA-256 checksum in Appendix A. Independent reinspection of that file reproduced the original and retained counts, found no missing cells, and verified the exact vector sensitivity mask. The checksum identifies the analyzed bytes and protects against silently substituting a different dataset edition.

Deduplication used every supplied column, including Time and Class. The retained rows were then stably sorted by Time. Removing equality duplicates can change class counts and partition membership, so raw dataset headline counts must not be used as denominators for the later model evaluation. The retained dataset has 283,726 transactions and 473 fraud labels.

The raw transactions are not redistributed with the supplement. Source links, attribution, download instructions, and the checksum support independent retrieval under the dataset's applicable terms. Public availability is not evidence of institutional representativeness, and no research ethics exemption is inferred from the access status.

B. REPRESENTATION AND INFORMATION AVAILABILITY

The supplied variables V1 through V28 are principal components. Their original financial meanings and upstream transformation fitting population are unavailable. Time and Amount are untransformed source fields. Time is used only to order and partition the observations; it is excluded from the model inputs. Amount receives a fixed logarithmic transformation.

$$x_{29} = \log(1 + \text{Amount})$$

Equation 1 Fixed transformation of the monetary amount input

Field group	Representation	Use in this study
V1 through V28	Supplied numeric principal components	Twenty eight predictors; original semantics unavailable
Amount	log of one plus the supplied amount	One additional predictor; no fitted transformation parameters
Time	Elapsed seconds from the source reference	Stable ordering and chronological split boundaries only
Class	Supplied genuine or fraud label	Supervised fitting and held out evaluation
Account identifiers	Not supplied	No account disjoint split or account clustered inference
Label confirmation time	Not supplied	No faithful delayed label availability simulation

Table 1: Input representation and permitted use

Corresponding email: saifulcu105@gmail.com

The resulting input matrix has 29 columns. Logistic regression standardizes them using means and scales fitted separately on that run's training subset. The tree, EBM, and histogram boosting receive the same unstandardized 29 input values. The logarithm requires no fitted parameters and does not use validation or test statistics.

A chronological split protects later observations from downstream fitting decisions, but it cannot certify the unknown PCA construction as leakage free. The study therefore claims controlled local preprocessing rather than a complete audit of the dataset's upstream representation. The distinction follows the learn and predict separation emphasized by Kaufman et al. (2012).

The label target is the source's transaction classification. It does not reveal customer intent, a verified compromise mechanism, recovery success, or the amount of loss prevented by an alert. These unavailable outcomes remain outside the measured claims.

C. CHRONOLOGICAL PARTITIONS AND LEAKAGE CONTROLS

Approximate 60, 20, and 20 percent row count cutoffs were adjusted to keep equal timestamp groups together. The retained observations have strictly ordered time ranges across the three partitions. Figure 3 and Table 2 give the actual split, which is the denominator for every reported result.

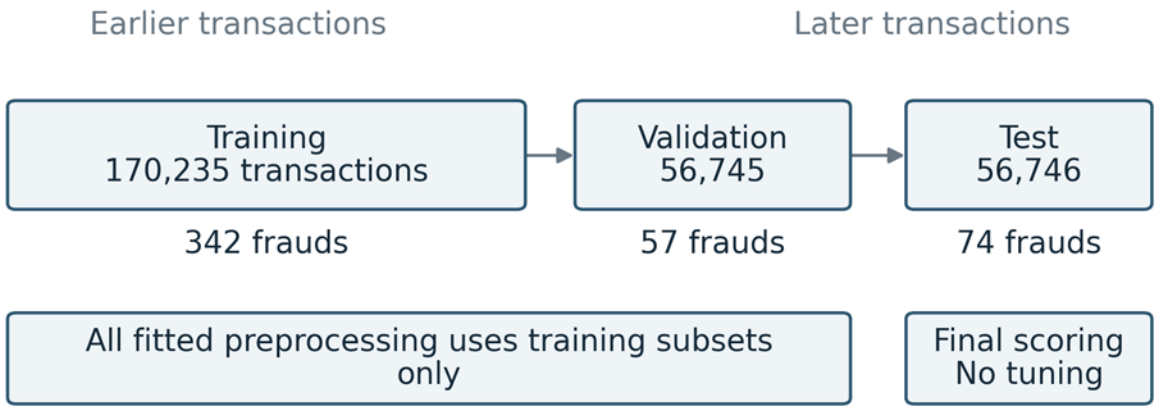


Fig 3: Timestamp preserving chronological evaluation. Model fitting uses the earlier training pool, threshold choice uses the following validation period, and final prediction uses the later test period. The boxes represent transaction counts rather than equal elapsed durations.

Partition	Transactions	Frauds	Fraud rate %	Time range s
Training	170,235	342	0.2009	0 to 120392
Validation	56,745	57	0.1004	120393 to 145233
Test	56,746	74	0.1304	145234 to 172792

Table 2: Verified chronological partitions

No model configuration, fitted scaler, selected threshold, or explanation reference cohort is chosen by maximizing test performance. The explanation diagnostic cases are selected by their labels to balance a diagnostic sample, but that sample is not used to train models or establish a population alert yield. Its selection is disclosed as a separate measurement decision.

Unknown label confirmation times prevent reconstructing when historical fraud judgments became available. A transaction's occurrence time is not a substitute for that clock. The benchmark is therefore a chronological classification experiment, not a fully point in time operational replay. Kapoor and Narayanan (2023) motivate making such information boundaries explicit rather than treating chronology alone as proof that all leakage is absent.

D. NESTED LABELED TRAINING BUDGETS

Five random seeds, 11, 29, 47, 71, and 97, each define one permutation of the fixed training pool. The first 10, 25, 50, and 100 percent of that permutation are selected, with integer truncation of the sample size. The selected indices are sorted before fitting. Every learner receives the identical subset for a particular seed and fraction.

$$n_f = \lfloor f n_{\text{train}} \rfloor, \quad \mathcal{I}_{s,f} = \{\pi_s(1), \ldots, \pi_s(n_f)\}$$

Corresponding email: saifulcu105@gmail.com

Equation 2 Seeded nested subsets of a fixed historical training pool

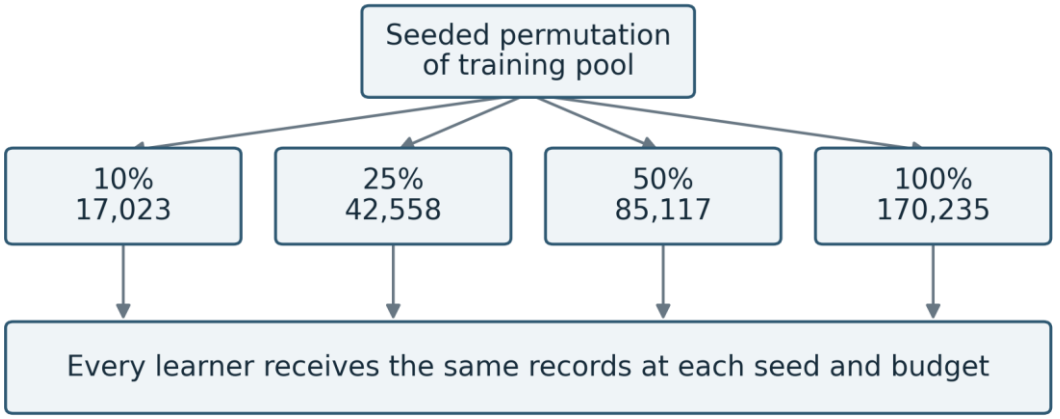


Fig 4: Matched nested training budgets. Each smaller subset is contained within the larger subset for its seed. The chronological training interval remains unchanged. Four learners share each subset, so differences do not arise from giving them different labeled records.

The resulting training sizes are 17,023, 42,558, 85,117, and 170,235 transactions. Sampling is not stratified, so the number of fraud examples varies naturally at the smaller budgets. At 10 percent, the five subsets contain 30, 34, 32, 37, and 20 frauds. At full data, every seed uses the same 342 frauds and the same complete training records.

This design approximates randomly missing historical labels. It does not model active learning, investigation selection, delayed feedback, an evolving collection period, or the economics of obtaining additional fraud labels. Small subsets can be difficult because both their total size and their observed rare class count change.

Independent checks reconstructed each seeded permutation and verified every saved training index array across all four models. This establishes the implemented pairing and nesting. It does not make the five model evaluations independent samples of future transaction populations, since all are evaluated on the same test cohort.

E. COMPARATIVE ARCHITECTURE AND FIXED CONFIGURATIONS

Four learners provide complementary controls: a linear model, a constrained single tree, an additive nonlinear model, and a boosted tree ensemble. Their configurations were fixed without a hyperparameter search. Eighty primary fits arise from four learners, four training fractions, and five seeds. A single full budget EBM interaction ablation adds one fit.

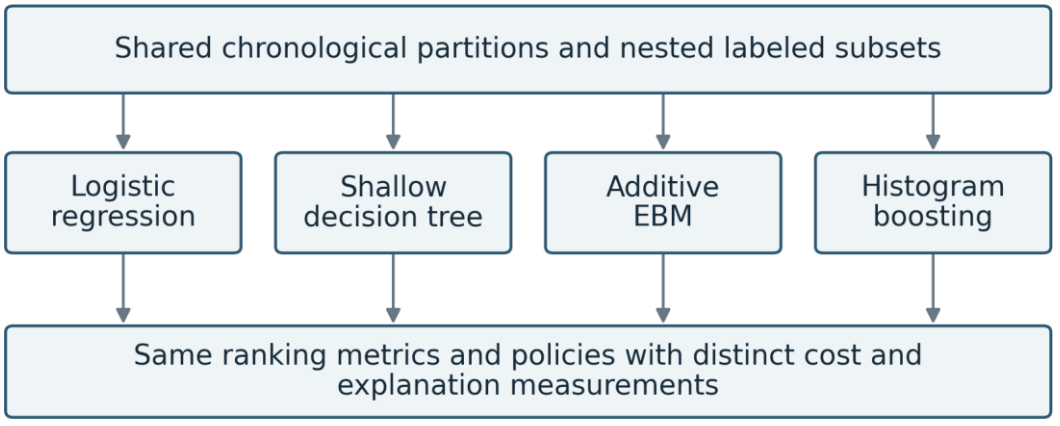


Fig 5: Controlled information exposure across learners. All learners use the same 29 inputs and matched training indices. The common evaluation does not imply equivalent architecture, capacity, computational effort, or explanation semantics

Corresponding email: saifulcu105@gmail.com

Learner	Main settings	Weighting
Logistic regression	StandardScaler; C 1; maximum 1,000 iterations	Balanced class weights
Shallow tree	Maximum depth 5; minimum leaf size 20	Balanced class weights
Additive EBM	No interactions; 4 outer bags; 64 bins; learning rate 0.05; 1,000 rounds	Balanced sample weights
Histogram boosting	150 iterations; maximum 15 leaves; learning rate 0.05; L2 value 1	Balanced class weights

Table 3: Explicit configuration choices

Scikit-learn 1.8.0 supplies logistic regression, the tree, and histogram boosting; InterpretML 0.7.8 supplies EBM. SHAP 0.54.1 provides the tree attribution calculations. Pedregosa et al. (2011) describe the scikit-learn framework. Exact package versions, remaining defaults, and execution details are retained because nominal model names alone are insufficient for reproduction.

The configurations encode different restrictions rather than equal parameter or training time allocations. The experiment therefore compares these configured learners. It cannot establish the best attainable performance of each family or rule out a better tuned shallow tree, regularized linear model, or EBM.

F. LOGISTIC REGRESSION AND BALANCED FITTING

Logistic regression combines a fitted standardization step with a regularized linear score. Its coefficient contributions are straightforward to compute in the standardized input space. They can expose the arithmetic behind a prediction while remaining limited by the anonymized meaning of the underlying variables.

$$w_c = \frac{n}{2n_c}, \qquad c \in \{0, 1\}$$

Equation 3 Balanced class weight calculated within each training subset

For class c, the subscripted count n_c is its frequency in the fitted subset and n is that subset's size. The weights are recalculated at every budget and seed rather than inherited from the full dataset. He and Garcia (2009) review the learning problems created by imbalance; weighting is one response to that problem and does not create additional observed fraud cases.

$$g(x) = \beta_0 + \sum_{j=1}^{29} \beta_j z_j, \quad p(x) = \frac{1}{1 + \exp[-g(x)]}$$

Equation 4 Linear raw score and logistic probability transformation

In Equation 4, z denotes the features standardized using that fit's training observations. The fitted scaler is then applied without modification to validation and test transactions. The regularization control is C equal to 1, and the optimizer has at most 1,000 iterations. The source code and software version define the remaining solver defaults.

$$\mathcal{L} = -\sum_i w_{y_i} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] + \lambda \|\beta\|_2^2$$

Equation 5 Schematic weighted regularized logistic objective

Equation 5 shows the role of class weights and coefficient regularization; library specific scaling determines the precise correspondence between lambda and C. The experimental setting is identified by the executed configuration rather than by an unreported hand chosen lambda. No feature selection or later test based coefficient adjustment is performed.

At full training data, the class one weight is approximately 248.88, while the class zero weight is approximately 0.501. At the smallest seed 97 subset, 20 fraud labels imply a class one weight of 425.575. These are objective weights, not calibrated risk multipliers or monetary misclassification costs.

The native explanation for feature j is its standardized value multiplied by the fitted coefficient. The sum of these terms and the intercept is compared with the raw model score. This is a numerical decomposition. Its correctness does not establish that a principal component corresponds to a comprehensible financial reason.

G. SHALLOW TREE SCORES AND PROBABILITY INTERPRETATION

Corresponding email: saifulcu105@gmail.com

The single tree baseline partitions the input space through a sequence of tests. Maximum depth five and minimum leaf size twenty constrain its complexity. A shallow architecture can be inexpensive to score, but it can also be an inadequate representation for a difficult class boundary. The comparison does not interpret weak performance under this cap as proof that all decision trees are unsuitable.

$$q(x) = \sum_{\ell=1}^{L} q_\ell \mathbf{1}\{x \in R_\ell\}$$

Equation 6 Piecewise constant tree output over learned leaf regions

In Equation 6, each region R is one leaf and q is its fitted class one score. Balanced class weights influence split selection and leaf class proportions. A model can therefore emit high weighted scores for cases whose natural population fraud probability is much lower. The term probability output should not be read as a calibration claim.

$$q_w(x) = \frac{w_1 p(x)}{w_1 p(x) + w_0[1 - p(x)]}$$

Equation 7 Idealized relationship between a weighted target and a natural class probability

Equation 7 describes the population optimum of a weighted binary log loss under simplifying assumptions. It illustrates why changing the class weights changes the target represented by an output q. It is not used to retrospectively recalibrate the reported predictions. Finite sample error, regularization, model misspecification, and distribution shift can prevent a simple formula from recovering reliable risks.

Dal Pozzolo et al. (2015) study calibration issues under undersampling. The present experiment uses weighting rather than that sampling procedure, so their work motivates careful probability interpretation without supplying an exact correction for these fitted models. Van Calster et al. (2019) likewise emphasize that calibration must be assessed separately from discrimination.

Tree SHAP is applied to the shallow classifier in its class one probability output space. This differs from the log odds space used for the logistic, additive EBM, and histogram boosting explanations. Consequently, contribution magnitudes are not compared across families. Even matching output scales would not make native and SHAP explanations identical estimands.

The fixed top five stability measure can include inactive tree terms if fewer than five contributions are nonzero. The study therefore retains both its declared fixed length result and an exploratory active only sensitivity. This distinction becomes material at the smallest tree training budget.

H. ADDITIVE EXPLAINABLE BOOSTING MACHINE

The additive EBM represents its raw score as an intercept plus learned univariate feature functions. Hastie and Tibshirani (1986) provide the generalized additive model foundation. Nori et al. (2019) describe the InterpretML implementation, while Lou et al. (2013) examine extensions containing a small number of pairwise interactions.

$$g(x) = \beta_0 + \sum_{j=1}^{29} f_j(x_j)$$

Equation 8 Main effects additive score used by the primary EBM

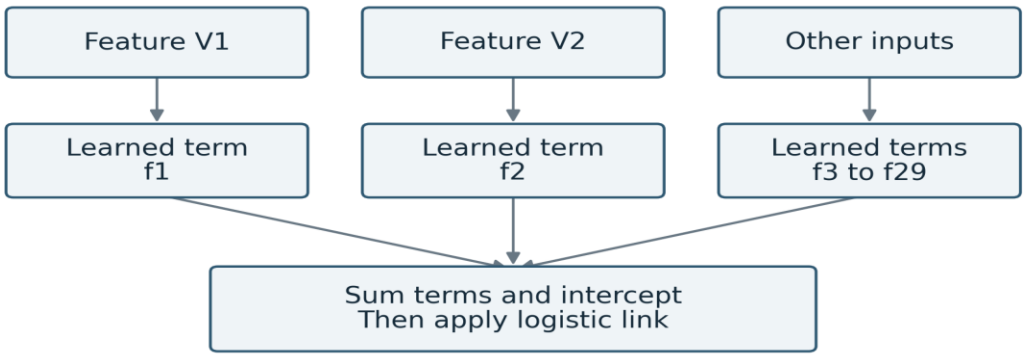


Fig 6: Additive EBM representation. Each input contributes through a learned univariate term, and the intercept and terms combine on the raw score scale. The schematic functions are architectural elements, not plotted explanations for a particular anonymized fraud case.

Corresponding email: saifulcu105@gmail.com

The main experiment sets interactions to zero, outer bags to four, maximum bins to 64, learning rate to 0.05, and maximum boosting rounds to 1,000. Early stopping is disabled. The package default internal validation fraction is retained for bag construction within the provided training subset; the external validation and test partitions are never supplied to this fitting procedure.

$$g_{\text{int}}(x) = \beta_0 + \sum_j f_j(x_j) + \sum_{(j,k)\in\mathcal{P}} f_{jk}(x_j, x_k)$$

Equation 9 Optional interaction representation evaluated in one diagnostic fit

The single ablation enables five EBM interactions at full data for seed 11. It is compared with the matched additive run. One seed cannot estimate the variability of interaction benefits, and the result is not used to replace the primary additive configuration after test inspection.

EBM explanations use the executed model's term contributions. The simple additive form makes a score decomposition accessible, but the learned functions still operate on principal components with undisclosed financial meanings. A native decomposition therefore supports model inspection without proving that an investigator can understand or act on the terms.

I. HISTOGRAM GRADIENT BOOSTING

Histogram gradient boosting combines many small trees through sequential updates to a raw score. The implementation bins numeric features and uses the loss information at the current score to construct subsequent trees. Friedman (2001) provides the gradient boosting foundation. The specific histogram implementation and its defaults are identified by the pinned scikit-learn version.

$$g_m(x) = g_{m-1}(x) + \eta h_m(x), \qquad m = 1, \ldots, 150$$

Equation 10 Additive tree updates under the fixed 150 iteration configuration

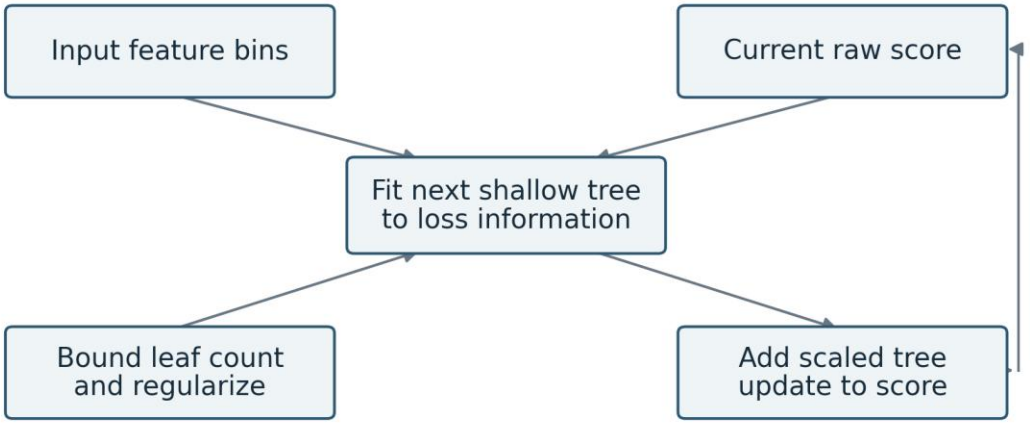


Fig 7: Simplified boosted tree update path. Feature bins, the current score, bounded tree complexity, and scaled updates determine the next score. The diagram omits implementation details and does not imply that any particular tree was newly designed for this study.

The experiment uses 150 iterations, maximum 15 leaf nodes, learning rate 0.05, L2 regularization value 1, balanced class weights, and no early stopping. These settings provide a strong tabular reference without a search over architectures. They do not match the tree baseline's capacity or the EBM's fitting budget.

Tree SHAP explanations are computed using the tree path dependent setting without an external background dataset. Histogram boosting contributions are evaluated on the raw score scale. Score reconstruction verifies the arithmetic at that scale; an explanation's numerical agreement is separate from a claim about why a transaction was actually fraudulent.

The comparison is particularly useful because training, scoring, and explaining an ensemble can have different costs. A short fit does not imply that every later workload is inexpensive. The measured prediction and explanation batches therefore remain separate endpoints rather than being combined into a general claim of speed.

J. RANKING ERRORS AND PROBABILITY MEASURES

Average precision is the primary ranking measure. It is calculated using scikit-learn average_precision_score and summarizes precision at successive increases in recall. It is not silently replaced by a trapezoidal area under a precision recall curve. Saito and Rehmsmeier (2015) motivate attention to precision and recall when the positive class is rare.

Corresponding email: saifulcu105@gmail.com

$$\mathrm{AP} = \sum_{k} (R_k - R_{k-1}) P_k$$

Equation 11 Average precision from changes in recall and corresponding precision

A constant score assigns the same rank to every transaction and has AP equal to test prevalence, 74 divided by 56,746, approximately 0.001304. This no skill reference matters when a numerical fit degenerates. High overall accuracy would be uninformative in a population containing so few positive examples.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

Equation 12 Threshold specific precision and recall

For an empty alert set, precision is undefined and remains missing. Recall is zero when no positive case is alerted. Confusion counts and alert rate are retained alongside ratios so that a strong looking percentage can be traced to its underlying numerator and denominator. Sokolova and Lapalme (2009) discuss how classification measures respond to different confusion matrix changes.

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2$$

Equation 13 Brier score for the emitted class one probabilities

ROC-AUC and Brier score are secondary measures. Gneiting and Raftery (2007) provide the proper scoring rule foundation for probability evaluation. The Brier score can be dominated by the many genuine transactions; a low aggregate value does not independently establish accurate minority class risk estimates. No post hoc probability calibration was performed.

Hand (2009) discusses limitations of interpreting area under the ROC curve as a universal performance criterion. The present experiment therefore reports that measure while keeping its operational comparisons at specified alert policies. A model with slightly lower aggregate AP can still have higher recall at one transported threshold, especially when its alert count differs.

The metrics answer different questions. Ranking addresses score ordering, threshold counts describe one action policy, and probability scores assess emitted numerical risks. None identifies prevented loss, successful intervention, or customer experience without additional outcome data.

K. FROZEN THRESHOLDS AND BATCH RANKING

Two policies are evaluated because a transported threshold and a capacity constrained ranked batch have different information requirements. The validation policy chooses a complete score cutoff that maximizes recovered frauds subject to alerting at most 0.5 percent of validation transactions. Precision resolves relevant ties; the executable routine is provided for exact reproduction.

$$\tau^* \in \underset{\tau:\ \widehat{\mathrm{AlertRate}}_v(\tau) \le 0.005}{\arg\max} \ \widehat{\mathrm{Recall}}_v(\tau)$$

Equation 14 Primary validation threshold selection objective

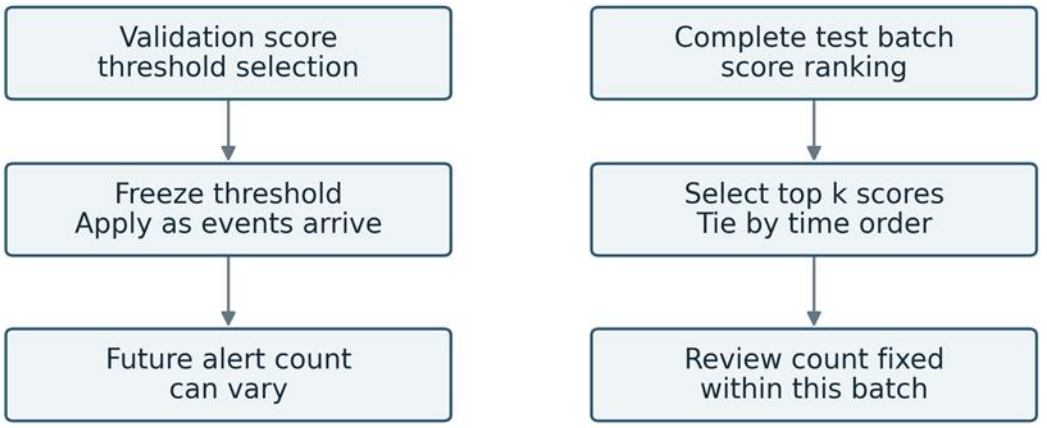


Fig 8: Distinct information requirements of the evaluated policies. A frozen threshold can score one arrival, while the ranking rule requires the complete test batch. The validation burden cap does not guarantee the same future test volume

Corresponding email: saifulcu105@gmail.com

$$k_b = \lceil b\, n_{\text{test}} \rceil, \qquad b \in \{0.001, 0.005, 0.01\}$$

Equation 15 Number reviewed under each retrospective test batch budget

The chosen threshold is frozen and applied using a greater than or equal comparison. Test alert rate is measured rather than forced to equal the validation cap. The separate ranking experiment selects the largest k scores in the complete test partition. Equal scores follow stable chronological row order. At 0.1, 0.5, and 1 percent, the reviewed counts are 57, 284, and 568.

The fixed ranking count makes a comparison at that batch workload clear, but it does not create a streaming policy. A live institution would need an interval such as a shift or day and rules for capacity reservation and late arrivals. No such operational scheduling was measured. The full test partition is the comparison population here.

Implementation detail is retained when it affects reproducibility. A constant validation score with no admissible nonempty tied group yields an infinite threshold and no alerts. For a remaining zero true positive candidate tie, the archived routine's ordering determines the selected cutoff; no claim about optimal customer burden is inferred from that fallback behavior.

L. COMPUTING WORKLOADS AND MEASUREMENT LIMITS

The original fits ran sequentially in separate Python interpreters on an AMD EPYC 9V74 cloud host with nine visible logical CPUs. Native numerical threading was capped at two, and EBM n_jobs was two. Training time measures the pipeline fit, including logistic standardization, while excluding dataset download and parent preparation. These are measurements on the recorded host, not hardware independent constants.

Quantity	Measured operation	Interpretation limit
Training time	Wall time around pipeline fit	Excludes download and parent preprocessing
Prediction batch	1,024 test rows after warm up; seven repeats	Median and 95th percentile describe this batch workload
Explanation batch	Same 100 diagnostic rows; seven repeats	Ongoing explanation work; setup recorded separately
Serialized size	Pickle byte length of the fitted model	Storage representation rather than runtime RAM
Process high water RSS	Maximum process reading for the workload	Common setup or inherited high water prevents model ranking
EBM worker memory	Not summed into a process tree profile	No total runtime memory comparison established

Table 4: Workload definitions and exclusions

Prediction timings use a warm up call followed by seven repeats on the same batch. The median is the main batch statistic, and the 95th percentile of those seven observations is retained descriptively. Seven repetitions do not establish a reliable production tail latency guarantee under concurrent load.

The explanation cohort comprises the 50 earliest test frauds and the 50 earliest genuine transactions, with their combined positions sorted chronologically. It is intentionally balanced to inspect model behavior. The resulting timing and stability summaries are not prevalence weighted estimates of explanation work across ordinary transactions.

Process high water RSS readings were identical across final fits. The telemetry cannot separate a fitted model's memory from shared preparation and accounting effects, and separate EBM worker memory was not summed. The values remain in the raw records but are excluded from comparative memory efficiency claims.

A local deployment would also include feature retrieval, policy execution, logging, concurrency, and failure handling. The present fit, prediction, and explanation batches do not measure that full service path. Serialized size is reported because it is observable, while its relationship to resident memory remains unmeasured.

M. EXPLANATION RECONSTRUCTION AND CONSISTENCY

Corresponding email: saifulcu105@gmail.com

The study evaluates two implementation properties on the fixed diagnostic cohort. Reconstruction asks whether contributions and a baseline reproduce a fitted model output. Stability asks whether the largest contribution sets persist across fitted models. Neither quantity measures whether an investigator understands the explanation or reaches a better decision.

$$\hat{g}(x) = \phi_0 + \sum_j \phi_j(x)$$

Equation 16 Reconstruction of a declared model output scale

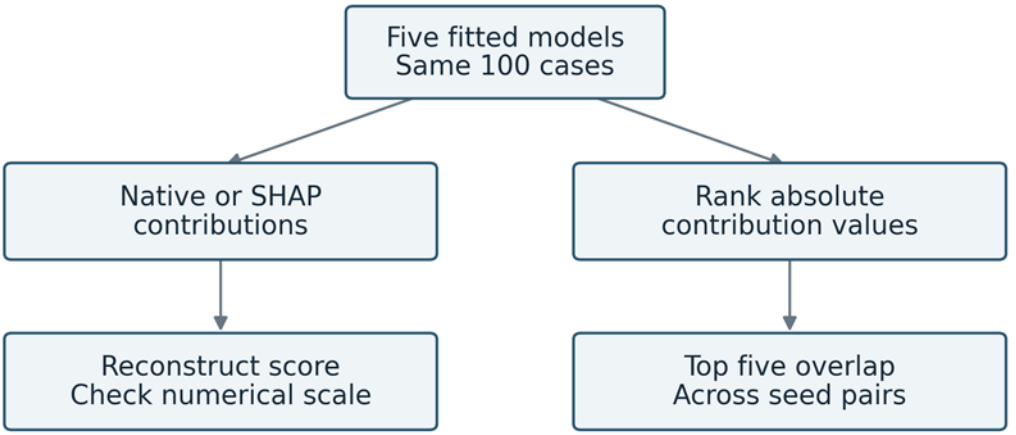


Fig 9: Separate explanation checks. Numerical reconstruction uses the model's specified output scale. Contribution overlap compares selected feature sets across ten seed pairs on the same 100 diagnostic cases. The two checks have different estimands and remain separate from human usefulness.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Equation 17 Jaccard overlap of two selected contribution sets

Logistic contributions are standardized input values multiplied by coefficients. EBM uses eval_terms. The tree and histogram ensemble use tree path dependent Tree SHAP. The shallow tree's attribution scale is class one probability; logistic, EBM, and histogram boosting use raw log odds. Absolute magnitudes are not compared across methods.

For each case and fit, the five largest absolute contributions are selected using stable feature order to resolve ties. Five seeds create ten model pairs, each evaluated on 100 cases, giving 1,000 pair and case comparisons per family and budget. Their mean overlap is descriptive. The comparisons share cases and fits and are not 1,000 independent experiments.

Lundberg et al. (2020) support tree based attribution methods, while Barredo Arrieta et al. (2020) emphasize that explanation purpose depends on the intended audience. The present properties concern a model's representation and numerical behavior. V1 through V28 cannot be converted into merchant, device, location, or customer narratives from the supplied documentation.

N. UNCERTAINTY AND SUPPLEMENTARY ANALYSES

Primary summaries are arithmetic means and sample standard deviations across five fits at a given learner and training budget. At smaller budgets, variation combines the composition of the sampled records with algorithmic behavior. At full data, all seeds share the same training records, so remaining variation is algorithmic or numerical rather than training sample uncertainty.

$$\bar{z} = \frac{1}{5}\sum_{s=1}^{5} z_s, \qquad s_z = \sqrt{\frac{1}{4}\sum_{s=1}^{5}(z_s - \bar{z})^2}$$

Equation 18 Descriptive across fit mean and sample standard deviation

Every fit is evaluated on the same later cohort containing 74 frauds. Five seeds do not create five independently observed fraud populations. A change of one recovered fraud moves recall by 100 divided by 74, or approximately 1.351 percentage points. Small observed differences require that granularity to remain visible.

Corresponding email: saifulcu105@gmail.com

Source	Varied here	Interpretation
Training records at partial budgets	Yes through five seeded permutations	Conditional sensitivity to sampled historical labels
Model algorithmic randomness	Yes where the learner uses its seed	Fit to fit repeatability under one protocol
Future test periods	No	No long horizon temporal uncertainty estimate
Independent institutions	No	No empirical transferability estimate
Account dependence	Identifiers unavailable	No defensible account clustered confidence interval
Human explanation use	Not studied	No trust or investigator effectiveness estimate

Table 5: Variation sources and what the experiment can estimate

Dietterich (1998) examines statistical comparisons of learning algorithms, and Brown et al. (2001) discuss uncertainty for binomial proportions. Their work underscores that an inferential procedure needs an appropriate sampling unit and assumptions. The present study does not convert its descriptive deviations into p values, population confidence intervals, or equivalence conclusions.

Two exploratory analyses are identified explicitly. One removes test predictor vectors exactly matching development period vectors, using saved predictions without refitting. The other excludes numerically inactive contributions from the top five sets. They diagnose sensitivity of the reported quantities and do not replace the primary analysis after favorable results are inspected.

RESULTS ACROSS LABELED TRAINING BUDGETS

The experiment completed all 80 primary fits and the one EBM interaction ablation. Recalculation from saved test probabilities reproduced AP, ROC-AUC, and Brier score for all 81 records exactly. Confusion counts, ranking policies, the nested indices, and explanation overlaps were independently checked. Figure 10 reports the principal learning curves.

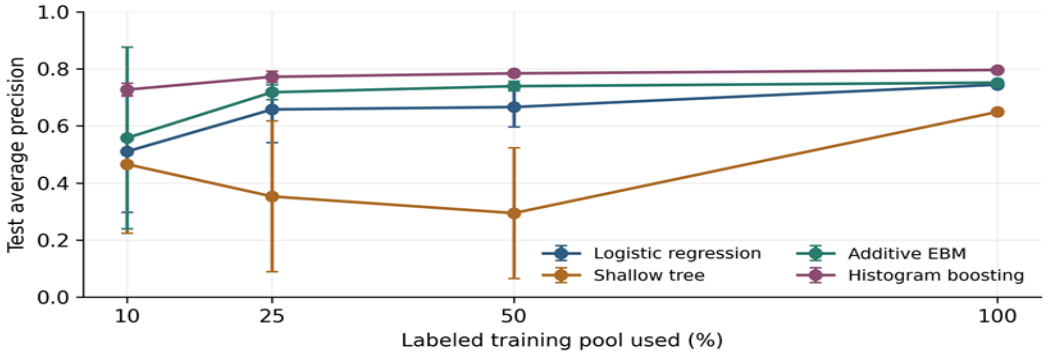


Fig 10: Test average precision under the four labeled training budgets. Error bars are sample standard deviations across five fits on the same fixed test period. They represent descriptive fit variation rather than confidence intervals for future transaction performance.

Model	10%	25%	50%	100%
Logistic regression	0.510 ± 0.213	0.657 ± 0.117	0.666 ± 0.070	0.7442 ± 0.0000
Shallow tree	0.466 ± 0.242	0.353 ± 0.264	0.294 ± 0.229	0.6488 ± 0.0002
Additive EBM	0.557 ± 0.319	0.718 ± 0.026	0.739 ± 0.017	0.7512 ± 0.0053
Histogram boosting	0.727 ± 0.023	0.772 ± 0.019	0.784 ± 0.004	0.7957 ± 0.0000

Table 6: Average precision as mean plus or minus sample standard deviation

At full data, histogram boosting has the highest observed mean AP, followed by additive EBM, logistic regression, and the shallow tree. This ordering concerns the tested configurations on one test period. It is not a significance finding or a claim that optimized alternatives would preserve the ranking.

The shallow tree curve is not monotonic. More fitted records under a fixed architecture do not guarantee higher measured AP on a particular later cohort. The configurations, their randomness, and class composition can interact. The study retains that behavior rather than smoothing the curve into an expected monotone trend.

Corresponding email: saifulcu105@gmail.com

A. RARE CLASS EXPOSURE AND APPARENT DATA EFFICIENCY

The label budget changes both total sample size and the number of observed frauds. Figure 11 shows the actual class one exposure by seed. This matters because an institution's label scarcity is often concentrated in confirmed rare cases rather than distributed evenly across all records.

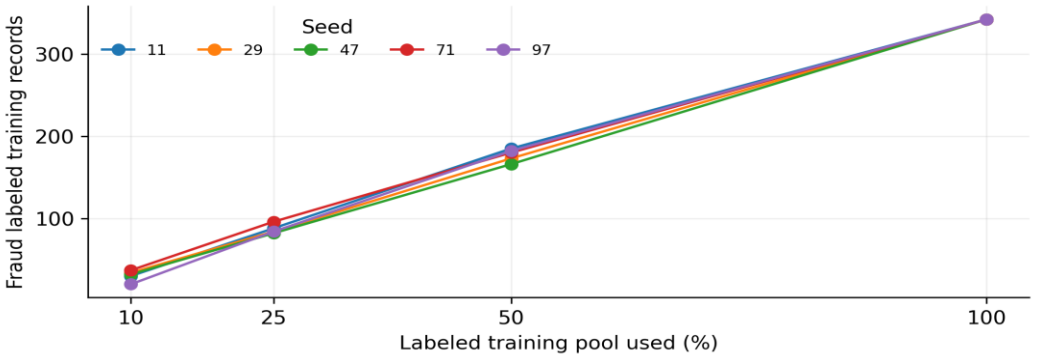


Fig 11: Fraud labeled training records in the nested subsets. The count varies across seeds at partial budgets and converges to 342 at full data. These are observed sample counts from the executed protocol, not estimates of an optimal investigation budget.

Seed	10%	25%	50%	100%
11	30	88	185	342
29	34	84	173	342
47	32	82	166	342
71	37	96	180	342
97	20	84	182	342

Table 7: Verified fraud counts by training fraction

$$\mathrm{RelativeAP}_{10} = \frac{\overline{\mathrm{AP}}_{10\%}}{\overline{\mathrm{AP}}_{100\%}}$$

Equation 19 Descriptive AP retention relative to a learner's full budget mean

Histogram boosting retains approximately 91.4 percent of its full budget AP at 10 percent training data. Additive EBM retains approximately 74.2 percent when every declared fit, including the constant score failure, is included. These ratios are descriptive normalizations. They do not establish minimum sufficient label counts, equivalent performance, or label acquisition savings.

The EBM failed subset has 20 frauds, but sample size alone is not an established cause of its numerical behavior. Other seeds have different records as well as different counts. A causal diagnosis would require a separate controlled investigation. The following failure analysis reports the observed outcome and its repeat without assigning an unsupported mechanism.

B. ALERT WORKLOAD AND THRESHOLD TRANSPORT

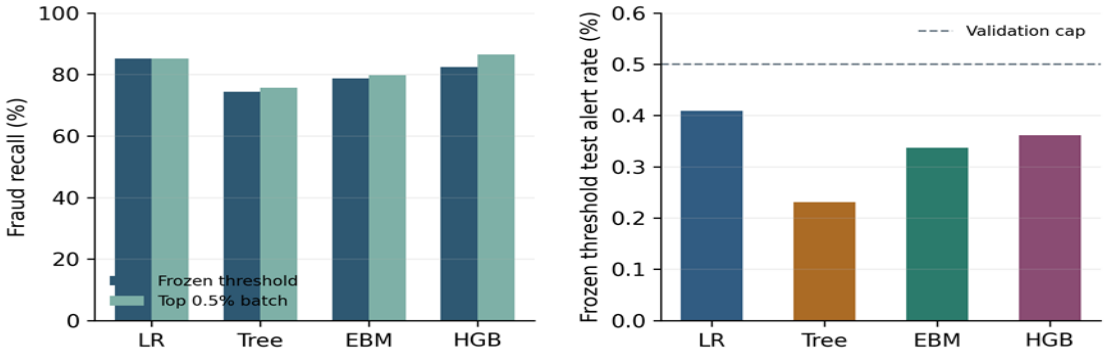


Fig 12: Full budget policy behavior. The left panel compares mean recall at the frozen validation threshold with complete test batch top 0.5 percent ranking. The right panel shows realized test alert rates for the frozen thresholds. LR, Tree, EBM, and HGB denote the four configured learners.

Corresponding email: saifulcu105@gmail.com

Model	ROC AUC	Brier	Top 0.5% recall	Frozen recall	Frozen alert %
Logistic	0.9815	0.02513	85.14%	85.14%	0.4088
Tree	0.8415	0.02468	75.68%	74.32%	0.2309
EBM	0.9628	0.00487	79.73%	78.65%	0.3373
HGB	0.9823	0.00168	86.49%	82.43%	0.3613

Table 8: Full budget test results under the two policies

At the top 0.5 percent ranking budget, exactly 284 test transactions are reviewed. Histogram boosting recovers 64 of 74 frauds, compared with 63 for logistic regression, 56 for the shallow tree, and a mean of 59 for additive EBM. The one fraud difference between the first two is approximately 1.35 recall points on this cohort.

The frozen threshold tells a different workload story. Logistic regression alerts 232 transactions and recovers 63 frauds; histogram boosting alerts 205 and recovers 61. Their recalls are therefore not a matched alert count comparison. Mean EBM counts can be fractional because five integer confusion matrices are averaged.

The Brier values describe the weighted, uncalibrated outputs. The all negative constant zero probability reference has Brier score equal to the test prevalence, approximately 0.001304. Its low aggregate error coexists with zero fraud detection at a positive threshold. This provides a concrete reason to retain recall, calibration context, and confusion counts alongside aggregate probability scores.

C. TRAINING PREDICTION AND EXPLANATION COSTS

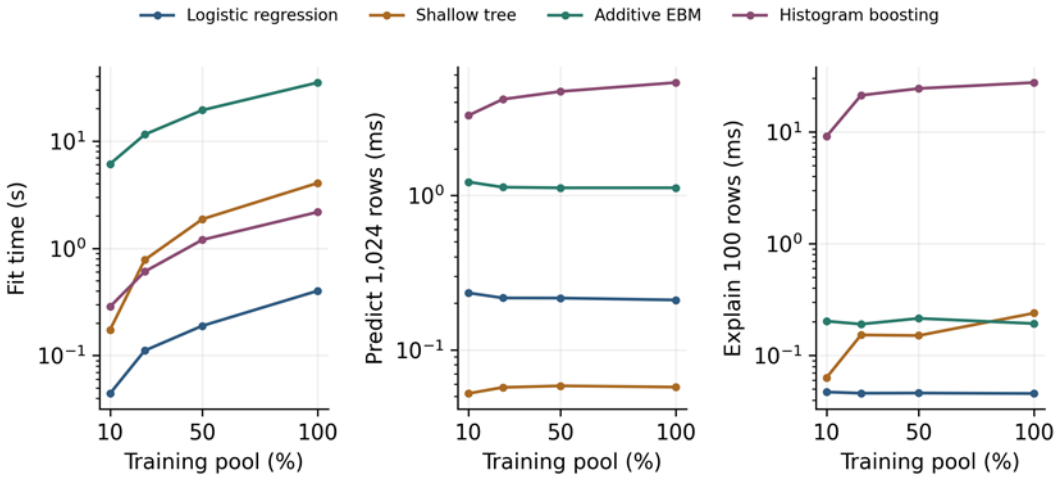


Fig 13: Recorded computing workloads across training budgets. Axes use logarithmic time scales. Each point is a mean of the five recorded fit or batch medians, as appropriate. The panels measure different operations and should not be combined into one service latency estimate.

Model	Fit s	Predict 1,024 ms	Explain 100 ms	Serialized KiB
Logistic	0.40	0.210	0.046	2.0
Tree	4.06	0.057	0.239	5.3
EBM	35.03	1.118	0.193	137.5
HGB	2.18	5.364	27.544	319.1

Table 9: Full budget resource summaries

Full budget EBM fitting takes a mean 35.03 seconds on the recorded host, compared with 2.18 seconds for histogram boosting, 4.06 for the shallow tree, and 0.40 for logistic regression. These differences measure the specified architectures and execution path. They are not family wide training cost constants.

Corresponding email: saifulcu105@gmail.com

Prediction and explanation rankings differ from fitting rankings. The shallow tree has the fastest measured 1,024 row prediction batch, while native logistic contribution calculation is the fastest measured explanation workload. Histogram boosting fits quickly in this comparison but has the largest recorded ongoing explanation time.

Serialized size ranges from approximately 2 KiB for logistic regression to 319 KiB for histogram boosting. That order describes the saved model representation. The flat process high water memory reading cannot substantiate a corresponding order in runtime RAM. A dedicated measurement of each process tree would be required for that claim.

Explanation setup and subsequent calls are distinct. The reported explanation column covers repeated calculations after the initial setup. An application that constructs a new explainer for every transaction could incur additional cost. The benchmark does not evaluate that service design or concurrent explanation demand.

D. WITHIN MODEL EXPLANATION STABILITY

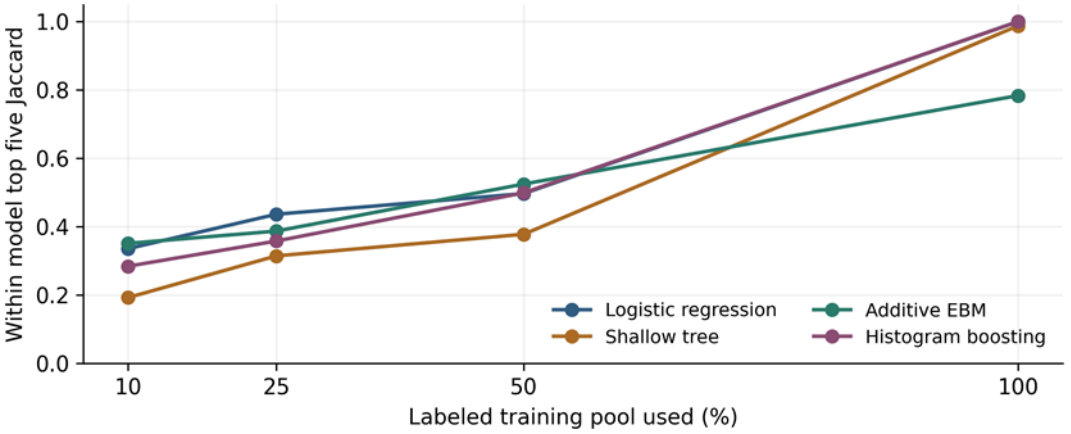


Fig 14: Mean top five Jaccard overlap across seed pairs and diagnostic cases. Every curve uses a within model comparison. The explanation methods, output scales, and fitting randomness differ, so the curves do not establish a common ranking of trustworthiness or human understanding.

Model	10% fixed top five	100% fixed top five	10% active only
Logistic regression	0.3356	1.0000	0.3356
Shallow tree	0.1925	0.9878	0.2250
Additive EBM	0.3511	0.7835	0.3511
Histogram boosting	0.2843	1.0000	0.2843

Table 10: Contribution set overlap at the endpoint budgets

All four families have lower observed set overlap at 10 percent than at full data. The result indicates sensitivity of selected contributions to the fitted training subset and algorithm under this diagnostic sample. It does not identify the explanation that would best help an investigator.

At full data, logistic regression and histogram boosting have overlap one because their fitted behavior is deterministic under the executed settings and shared records. Perfect repeated selection on the same cases is not certainty about a future population. EBM retains seed related variation through its fitting procedure, and the tree can change among tied choices.

At the 10 percent tree budget, 20 percent of case and fit combinations have fewer than five contributions larger than 1e-12. The fixed set therefore includes inactive terms chosen by feature order. Excluding inactive selections raises the exploratory mean overlap from 0.1925 to 0.2250. The primary fixed length result is retained together with that limitation.

The 1,000 overlap comparisons share both diagnostic cases and model fits. Their spread is a descriptive distribution over related comparisons rather than an independent sampling distribution for statistical inference.

E. RETAINED NUMERICAL FAILURE AT LOW DATA

Corresponding email: saifulcu105@gmail.com

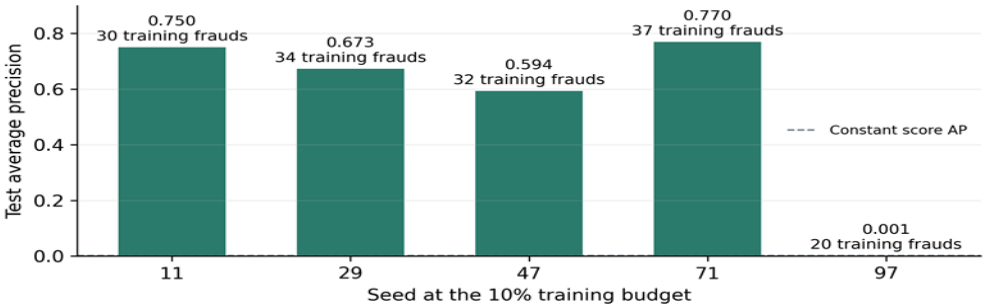


Fig15: All five additive EBM fits at the 10 percent budget. Seed 97 emits one constant zero test probability and has AP equal to prevalence. The annotation gives each subset's fraud count. The failed outcome remains in the primary mean and is not replaced by a selected successful run.

Seed	Training frauds	Test AP	Unique test scores
11	30	0.7497	55,403
29	34	0.6731	55,399
47	32	0.5936	55,382
71	37	0.7696	55,312
97	20	0.001304	1

Table 11: Low budget EBM diagnostic results

$$\varepsilon_{\text{abs}} = |\hat{g} - g|, \qquad \varepsilon_{\text{rel}} = \frac{|\hat{g} - g|}{\max(1, |g|)}$$

Equation 20 Absolute and scale relative reconstruction error

The archived exact diagnostic repeat reproduces the seed 97 test scores. Its intercept is approximately minus 7.956 times 10 to the power 13. The maximum absolute reconstruction difference is 0.078125, while the maximum relative difference is approximately 9.82 times 10 to the power minus 16. The nondegenerate fits have a largest absolute reconstruction difference of approximately 1.30 times 10 to the power minus 12.

A small relative arithmetic error at an enormous raw score scale does not validate the probabilities or explanations. All seed 97 test probabilities are zero, ROC-AUC is 0.5, and its frozen policy produces no alerts. The numerical scale is important to interpretation, but no causal mechanism involving weighting, bagging, or optimization was established.

Including all five fits gives mean AP 0.5575. Conditioning on the four nonconstant fits gives 0.6965, a different quantity. The latter is reported only to show the consequence of exclusion. It is not the primary result or evidence that a deployment rejection and fallback procedure has been tested.

F. INTERACTION AND REPEATED VECTOR SENSITIVITIES

The prespecified interaction ablation enables five EBM interactions at full training data for seed 11. It achieves test AP approximately 0.741, compared with 0.756 for the matched additive run. The single comparison is a diagnostic configuration sensitivity, not an estimate that interactions are generally harmful.

Analysis	Observed result	Permitted interpretation
One EBM interaction ablation	AP about 0.741 versus 0.756 for its matched main effects run	A single configuration sensitivity requiring further replication
Exact test vector matching	1,720 test rows match development predictors; zero frauds removed	Equality of supplied predictors rather than proof of shared account identity
Reduced evaluation cohort	55,026 transactions and all 74 test frauds	Changed evaluation population without model retraining
Tree active only overlap	0.2250 at the 10% budget versus fixed overlap 0.1925	Sensitivity of selected set definition to inactive contributions

Corresponding email: saifulcu105@gmail.com

Table 12: Supplementary analysis scope and observed outcomes

The exploratory vector analysis uses exact equality on V1 through V28 and the original Amount field, excluding Time and Class. Independent reinspection of the raw CSV reproduces the saved membership mask. Applying the fixed monotone amount transformation does not provide a business identity, and matching predictors alone cannot establish that two rows describe the same account or physical transaction.

Removing the 1,720 matched genuine rows leaves 55,026 test transactions and the original 74 frauds. Recalculation from saved predictions gives full budget AP values that round to 0.744 for logistic regression, 0.649 for the tree, 0.751 for EBM, and 0.796 for histogram boosting. The broad rounded ordering remains unchanged.

This sensitivity changes the cohort without retraining or modifying model choices. It cannot isolate a causal contamination effect, establish account independence, or repair unknown upstream PCA fitting. The analysis is retained as an exploratory audit rather than a claim that the primary dataset is independent after exclusion.

Both sensitivities have deliberately narrow interpretations. More seeds would be required to characterize interactions, while a semantically documented dataset with identifiers would be required to evaluate account duplication and related events directly.

G. INTERPRETING RESOURCE AND RANKING TRADEOFFS

Figure 16 combines two measured quantities at full training data: average precision and fitting time. Marker area encodes serialized representation size. It makes a limited tradeoff visible without turning storage into a runtime memory estimate or treating these three quantities as a complete utility function.

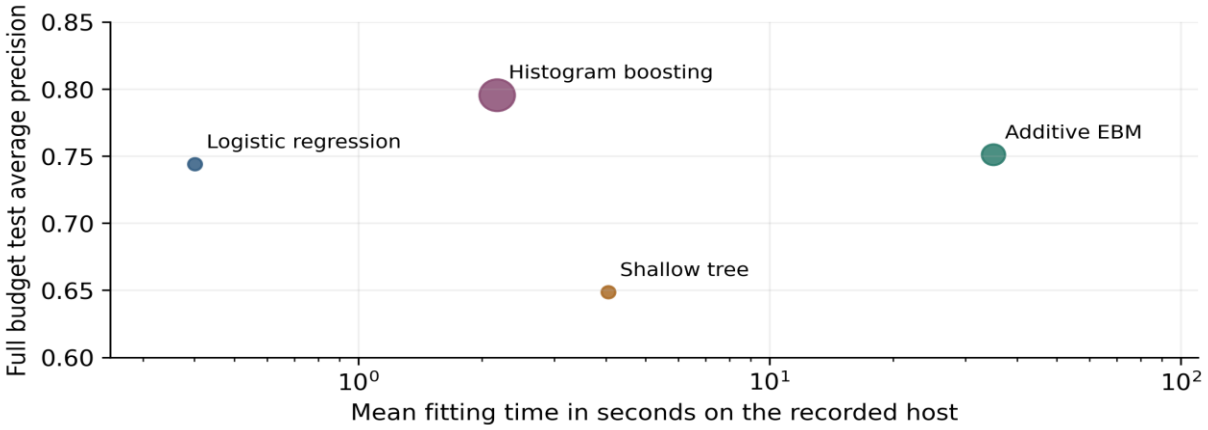


Fig16: Full budget mean AP against recorded fitting time. Point area is proportional to serialized model bytes with a small minimum marker size for visibility. Neither unmeasured investigator usefulness nor full service latency is represented. This is a descriptive comparison of the configured learners.

$$\mathrm{BatchTimePerRow} = \frac{T_{1024}}{1024}, \qquad \mathrm{KiB} = \frac{\mathrm{SerializedBytes}}{1024}$$

Equation 21 Batch normalization and binary storage unit definitions

Dividing a batch time by 1,024 gives an amortized time per row for that batch. It does not estimate the latency of an isolated online request. For example, the histogram boosting median batch value of 5.364 milliseconds corresponds to about 5.24 microseconds per row under this arithmetic, while its single request overhead and feature retrieval remain unmeasured.

Histogram boosting has both higher AP and shorter fitting time than the configured EBM in this experiment. That observation does not settle the choice when native inspectability, prediction workload, explanation demand, maintenance, or a local memory limit matters. Logistic regression has a small serialized representation and very short measured fit, while the shallow tree offers a fast prediction batch under a stronger expressiveness cap.

A useful local comparison would specify hard limits and then report the detection performance of every model satisfying them. Applying thresholds selected after seeing these results would be an exploratory decision analysis. The present study therefore provides the measured quantities without inventing an institution's staffing costs, hardware budget, or preference weights.

DISCUSSION AND LIMITATIONS

Corresponding email: saifulcu105@gmail.com

The benchmark separates four questions that are often conflated. The learning curve concerns randomly reduced historical labels. The prediction and explanation measurements concern different workloads. Score reconstruction concerns implementation arithmetic. Contribution overlap concerns repeatability of selected terms. No single result validates the other three.

The principal scientific limitation is the source: one familiar anonymized dataset from 2013 covering approximately two days. Chronological evaluation supports a short horizon comparison but cannot establish robustness to seasons, economic changes, evolving customers, or sustained adversarial adaptation. Zuberi et al. (2026) make a similar distinction between short horizon variation and longer production drift.

The test period contains only 74 frauds, and account identifiers are absent. Repeated seeds provide evidence of fit behavior conditional on the held out transactions, not a larger independent event sample. Unknown label confirmation times prevent a faithful delayed feedback study. Matching exact predictor vectors cannot recover unavailable account or campaign relationships.

Upstream PCA provenance and feature meanings are incomplete. The local pipeline avoids fitting scalers or choosing configurations on the test set, but it cannot verify how the supplied components were learned. A contribution to V14 describes an input to the model; it cannot be translated into a merchant, device, or customer cause without external documentation.

The configurations were fixed rather than optimized under equal computational budgets. This improves protocol clarity while limiting family wide conclusions. A learner can be weak because the tested cap is unsuitable, and a larger tuning exercise could change the ranking. Comparisons with other papers are particularly sensitive to different splits, feature engineering, sampling, and threshold policies.

Probability calibration, investigator understanding, customer experience, fairness across customer groups, and realized financial losses were not measured. The source lacks the variables or study design needed to establish those outcomes. Weighted probability outputs and low Brier values are not substitutes for a local calibration assessment.

Resource evidence also has a boundary. Timings describe one cloud execution environment and defined batches. The memory telemetry does not permit a model RAM ranking, and serialized size does not resolve that absence. End to end request latency, concurrency, and resilience to service failures require a separate intended infrastructure evaluation.

The constant score EBM fit remains a material result rather than an outlier silently removed from the study. A future training system could reject such a fit and fall back to another learner, but that complete rejection and replacement policy would need its own evaluation.

A. PROSPECTIVE INSTITUTIONAL REPLICATION

A meaningful replication starts with an operational decision: what is being alerted, who reviews it, how quickly a response must occur, and which errors impose costs. The institution should define its available labels, acceptable intervention burden, intended explanation user, and response policy before comparing model families.

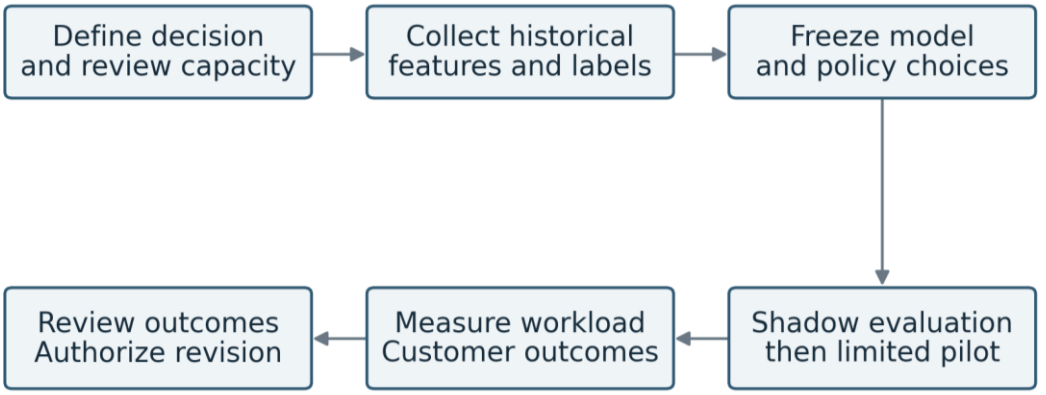


Fig 17: Proposed institutional replication path. Local definitions, longer records, frozen choices, prospective evaluation, and mature outcomes form a future study. These steps have not been completed at a financial institution as part of the current benchmark.

The data record should include stable entity identifiers, decision timestamps, feature availability, label confirmation, action execution, and outcomes. A longer chronological period permits evaluation across changing conditions. Gama et al. (2014) provide a foundation for studying concept drift, while Carcillo et al. (2018) emphasize feedback latency in fraud systems. Neither issue can be reconstructed reliably from a two day table without the relevant clocks.

Corresponding email: saifulcu105@gmail.com

A prospective shadow period should follow frozen model, calibration, and policy selection. Established institutional defenses can continue while recommendations are logged. Any later pilot would measure review burden, challenge completion, false intervention, recovery, and confirmed outcomes separately from the classifier's ranking score.

Explanation evaluation should use semantically documented variables and a declared task, such as identifying erroneous inputs or completing a case review accurately. A human study is required before claiming improved investigator decisions. Native model structure can help present an explanation, but it does not supply the missing effectiveness measurement.

The local study should profile process tree memory, full feature retrieval, scoring, explanation setup, logging, and concurrent load on intended hardware. Those quantities can change the preferred model even when the benchmark AP ordering remains similar. Proposed requirements are hypotheses to measure, not additional findings of the present experiment.

B. IMPLICATIONS FOR MODEL SELECTION

The strongest observed AP configuration is histogram boosting, but the broader conclusion is conditional model selection. An institution with infrequent retraining and substantial explanation demand can face a different tradeoff from one that repeatedly refits models but explains only a small share of alerts. The measured quantities should enter the actual decision at their relevant frequencies.

An additive model can expose a raw score decomposition directly while requiring longer fitting under the recorded settings. A tree ensemble can fit quickly and rank well while imposing a larger ongoing attribution workload. A linear model can offer inexpensive fitting and a small representation while remaining limited by its decision surface. These are properties of the tested configurations rather than universal rankings of interpretability or usefulness.

Rudin (2019) motivates considering inherently interpretable models rather than assuming that explanations can repair every opaque decision system. The present empirical results add a practical qualification: a claim of interpretability must be connected to input semantics and the user's task. A transparent equation over anonymous components is not automatically a meaningful explanation for a fraud investigator.

Failure retention changes the selection question. The mean AP over all declared EBM low budget fits is 0.5575; the mean conditional on nonconstant scores is 0.6965. Choosing the second value as the headline would evaluate a selected success condition. A real fallback policy may be sensible, but its failure rate, retraining cost, replacement performance, and total workload must be counted.

Recent work also places the contribution in context. Fazel et al. (2026) focus on optimized EBM configurations, while Gamal et al. (2026) examine broader hybrid architectures. Their findings cannot be treated as interchangeable comparisons because the protocols differ. The present study provides a compact fixed configuration baseline with traceable outputs, not a claim of state of the art performance.

External validation would be the most informative extension. A longer dataset with original variable meanings and independent customer groups could reveal whether the ranking and explanation tradeoffs persist. Repeating more seeds on the same two day test cohort would improve some fit stability summaries while leaving that central transfer question unanswered.

CONCLUSION

A reproducible experiment compared logistic regression, a shallow decision tree, additive EBM, and histogram gradient boosting across four historical labeled training budgets and five seeds. The study completed 80 primary fits and one interaction ablation. Independent recomputation verified every saved prediction record's AP, ROC-AUC, and Brier score and checked confusion counts, ranking outcomes, subset pairing, and contribution overlap.

On the full training pool, mean test AP was 0.7957 for histogram boosting, 0.7512 for additive EBM, 0.7442 for logistic regression, and 0.6488 for the shallow tree. Under top 0.5 percent complete batch ranking, histogram boosting detected 64 of 74 test frauds. Fixed validation thresholds produced different future volumes, making threshold transport and batch capacity distinct results.

Computing and explanation findings qualify a simple AP ranking. Mean full budget fitting times ranged from about 0.40 to 35.03 seconds on the recorded host, while prediction and explanation costs followed different orders. Full data deterministic fits had perfectly repeatable selected contribution sets on the diagnostic cases, but that repeatability does not establish causal meaning or human comprehension.

The retained low budget EBM failure demonstrates why complete reporting matters. A constant zero score has AP and Brier equal to prevalence, and its apparent numerical reconstruction behavior must be interpreted at its extreme raw score scale. Excluding the fit would change the estimand. A tested deployment pipeline must count failures and fallback behavior as part of the system.

Corresponding email: saifulcu105@gmail.com

The findings support conditional conclusions about the specified benchmark and configurations. They do not establish equivalence, long horizon resilience, real institutional resource savings, investigator effectiveness, or prevented financial loss. Those claims require a longer and semantically documented record together with locally measured review and infrastructure requirements.

REFERENCE

1. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
2. Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. Statistical Science, 16(2), 101–133. https://doi.org/10.1214/ss/1009213286
3. Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with spark. Information Fusion, 41, 182–194. https://doi.org/10.1016/j.inffus.2017.09.005
4. Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences, 557, 317–331. https://doi.org/10.1016/j.ins.2019.05.042
5. Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating Probability with Undersampling for Unbalanced Classification. 2015 IEEE Symposium Series on Computational Intelligence, 159–166. https://doi.org/10.1109/ssci.2015.33
6. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797. https://doi.org/10.1109/tnnls.2017.2736643
7. Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, 10(7), 1895–1923. https://doi.org/10.1162/089976698300017197
8. Fazel, R. E., Bakhtiary, A., & Bigdeli, S. A. (2026). Improving Credit Card Fraud Detection with an Optimized Explainable Boosting Machine. arXiv preprint arXiv:2602.06955. https://doi.org/10.48550/arXiv.2602.06955
9. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
10. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44, 1–37. https://doi.org/10.1145/2523813
11. Gamal, N., Younis, E. M. G., & Makram, W. M. (2026). Enhancing credit card fraud detection with a hybrid approach using machine and deep learning. Scientific Reports, 16(1), Article 10944. https://doi.org/10.1038/s41598-026-42891-4
12. Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102(477), 359–378. https://doi.org/10.1198/016214506000001437
13. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? Advances in Neural Information Processing Systems, 35. https://doi.org/10.52202/068431-0037
14. Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning, 77(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5
15. Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. Statistical Science, 1(3), 297–310. https://doi.org/10.1214/ss/1177013604
16. He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/tkde.2008.239
17. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804
18. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), Article 15, 1–21. https://doi.org/10.1145/2382577.2382579
19. Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 623–631. https://doi.org/10.1145/2487575.2487579

Corresponding email: saifulcu105@gmail.com

20. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
21. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
22. Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. arXiv preprint arXiv:1909.09223. https://doi.org/10.48550/arXiv.1909.09223
23. OpenML. (n.d.). creditcard dataset 1597 version 1. Dataset. https://www.openml.org/search?type=data&status=active&id=1597
24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html
25. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
26. Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
27. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
28. ULB Machine Learning Group. (n.d.). Credit Card Fraud Detection. Kaggle dataset. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
29. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. BMC Medicine, 17(1), Article 230. https://doi.org/10.1186/s12916-019-1466-7
30. Zuberi, N., Rahman, M. M., Rashid, S. U., Ahsan, M. N., Rahman, M. S., & Mollah, M. H. O. R. (2026). A robust machine learning framework for detecting temporal drift in financial fraud prevention. Scientific Reports, 16(1), Article 27776. https://doi.org/10.1038/s41598-026-58285-5

Corresponding email: saifulcu105@gmail.com