Research Article

Prior-Invariant Max-Hellinger Forests For Skew-Insensitive Classification of Rare Events

Jageti Padmavathi[1] , Vijay Kumar Burugari[2]

[1]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[2]Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

Abstract

Severely imbalanced binary learning breaks down when class counts, boundary overlap, label corruption and probability error are treated as a single problem. This paper addresses the thesis objective of designing a Max-Hellinger node criterion, demonstrating its class-prior invariance for fixed branch-conditionals, and characterizing a bounded bagged forest constructed from this criterion. The PI-MHF framework incorporates class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Three algorithms define the architecture, and three propositions characterize bounded or invariant behavior under explicit finite-support assumptions. A seeded controlled experiment employs 10,000-observation Gaussian-mixture profiles at imbalance ratios of 10, 30, and 100, increasing overlap, training-label corruption (as recorded), clean held-out targets, and repeated seeds. PI-MHF achieves minority recall 0.746, MCC 0.374, and average precision 0.486 in the severe profile, and 0.531, 0.220, and 0.261 in the extreme profile. Comparator methods are superior on some endpoints, so the evidence supports a conditional mechanism claim, not a claim of universal superiority. The results show when the objective mechanism is helpful, where its safeguards retain information, and which failure modes need to be evaluated with recorded data.

Index Terms—class imbalance, rare-event classification, ensemble learning, PI-MHF, probabilistic calibration, controlled. simulation.

I. INTRODUCTION

Class imbalance alters the link between empirical error and operational risk: a classifier makes predictions of dominant class for almost all observations, and yet it fails to catch events that have led to its deployment. Reviews of imbalanced learning report on this breakdown of performance on data, algorithm and ensemble levels [1], [2]. This is relevant to PI-MHF for the model needs to distinguish between class scarcity and observation reliability. A frequency re-weightment can raise the numerical weight of a rare class, but it should not infer if one labeled observation is an interior example, a valid boundary case or a corrupted record. The proposed mechanism delegates this choice to explicit statistics rather than an implicit preprocessing decision. Its output is bounded, logged, and assessed on partitions fixed in advance of any transformation. This system connects the mathematical object to an observable diagnostic, and therefore prevents the model description from claiming the existence of gains that the artificially controlled evidence doesn't display. The corresponding removal or parameter ablation quantifies the effect of this mechanism alone against the rest of the ensemble.

The hardness is determined not by the class ratio. It is currently unknown how small disjuncts, class overlap, label errors, and changes in sampling prevalence affect the learning process, and these effects might be worse than a moderate count difference [3], [4]. The question for the learner is not if imbalance should be considered, but rather how a learner should respond to the presence of rarity, overlap, noise, and finite support all at once. In PI-MHF, the answer is elaborated in the definition of class-conditional branch normalization, selection of Max-Hellinger splits, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Every value has a limited range and a documented fallback. That makes the analysis falsifiable: class-sensitive performance, probability quality, retained support, and computation can all be studied separately," he said. It also precludes interpreting overall accuracy as evidence of rare-event skill. The procedure is permitted to lose to a comparator on a single endpoint; that loss becomes part of the mechanism analysis rather than being obscured by a composite average.

Contemporary taxonomies still see resampling, cost-sensitive learning, calibrated decisions and ensemble construction as separate points in the intervention space [5]. Although SMOTE is still the reference data-level method, interpolation can replicate local overlap when minority neighborhoods are polluted[6]. A straightforward

consequence is then that our proposed stage cannot be judged only from the output class label. Intermediates values indicate the extent to which a result is driven by useful geometric change versus destructive drift of the effective class prior. The evaluation reports these quantities, associates them with minority recall, MCC, stability of the prior-responses, disagreement within the tree, convergence as a function of the size of the ensemble, and computational cost, and stores the seed and partition identifiers needed for repetition. This belief structure is at the heart for the boundary argument of the paper. It makes claims about the behavior of the declared generator and it does not turn the simulated rows into clinical, financial or operational observations. The same distinction applies to the interpretation of run-time, which should be viewed as a metric measured from the (emulated) execution environment, not as a transferable latency guarantee.

Boundary active synthesis was proposed with the effect of focusing the synthesis of samples near hard minority areas [7], and adaptive synthesis is a variable number of synthetic samples generated for each minority sample based on its local difficulty [8]. These procedures allow rarer patterns to be seen more often, but their action is based on a trust in the local neighbours. This design can be thought of as constrained estimation. The estimator is given an imbalanced labeled partition and returns both a classifier and an audit state. That the classifier is useful only when its class- sensitive utility is still high; the audit state is useful if a later reader can figure out why samples, splits, weights, or thresholds were accepted. PI-MHF satisfies these conditions by finite candidate sets, fixed stopping rules and numerical clipping. Theoretical results establish boundedness or concentration under given assumptions. They do not guarantee global optimality for the nonconvex forest, universal superiority, or transfer to an ever unseen domain.

Comparative evidence suggests that resampling results depend on the geometry of the dataset and the choice of learner [9]. Unilateral selection removes redundant majority examples and some boundary pairs, reducing the size of the training set at the cost of some information retention [10]. This formulation helps to reveal the relationship between the research gap and the research objective. Impurity reductions, computed from joint node counts, contract as the minority prior becomes small, hiding branch-conditionals that still are discriminative. The goal is to create a Max-Hellinger node criterion, demonstrate that it is unaffected by class-prior changes (assuming fixed branch-conditionals), and then investigate a bounded bagged forest constructed using this criterion. A proper test should control for class weightings, threshold tuning, and ensemble size, and isolate the statistic of interest. The experimental design provides those controls and shows the observed ordering. Finally, sensitivity analysis is performed, varying one mechanism parameter at a time, while the data generator, split, and metric definitions remain constant. This structure allows positive and negative findings to be interpreted within bounds and provides a clear framework for subsequent studies using recorded data to replicate.

Tomek links are a method of identifying and removing closely located opposite-class pairs [11], and edited nearest-neighbor rulesemit observations which are not in line with local votes [12]. In extreme scarcity, the same rules might also remove legitimate minority edges unless deletion is protected by support evidence. The corresponding protection has scientific merit as well as software merit. Without a finite guard, an observation whose local support is vanishing (or a node that contains very few points from the rare-class) can dominate a ratio, distance, or a posterior. A high numerical output may be misinterpreted as strong evidence. The PI-MHF avoids that breakdown via the stage-specific bounds provided in the protocol and summarized in the safeguards table. The proofs are the same bounds. The experiments measure their cost in terms of retention, calibration, disagreement, sensitivity, or fit time, so the safeguard is not a hidden implementation detail.

Recruiting information-spreaders via exploratory undersampling generates multiple majority subsets rather than relying on one removal pattern [13]. RUSBoost combines random undersampling with boosting [14], decreasing the amount of training but adding sampling variance. The link between mechanism and evidence holds conditional. The results vary depending on the specified Gaussian-mixture structure, the level of imbalance, the amount of overlap, the noise model, and the set of bounded learners. These are convenient conditions for studying internal mechanisms, since the ground truth and the corruption labels are available. They're not a replacement for a public benchmark rerun. The paper keeps this distinction close to each table and figure. And any subsequent employment (using healthcare, fraud, intrust, or edge data, though not these datasets) must refit preprocessing within the training partition, respect entity or temporal groupings, and evaluate domain-specific costs before issuing an operational verdict.

Cost-sensitive boosting modifies how much of the loss will be contributed by the costly errors [15]. The decision-theoretic basis requires costs or defensible surrogates of costs, which are frequently unavailable when only class counts are known [16]. From a decision analysis point of view, false negatives and false positives have different

interpretations. A higher rare-class recall can be bought at an unusable alert burden, and a high precision can be the result of rejecting most rare cases. the paper, therefore, interprets recall next to precision, MCC, average precision, balanced-accuracy and calibration. PI-MHF is evaluated by the shape of this trade-off and not by any single number. The design puts threshold determination on top of model fitting and confines it to the calibration data, thereby isolating the test set from every model and operating-point selection only the test set.

Bagging uses bootstrap diversity to reduce variance [17], and random forests bootstrap samples and randomizes feature search [18]. Bootstrap diversity and randomized feature search do not specify how rare or uncertain observations should influence training. This is an important consideration for PI-MHF, as the model needs to distinguish between class scarcity and observation reliability. While frequency correction can increase the numerical contribution of a rare class, it cannot determine whether a single labeled observation is an interior example, a true boundary case, or a noisy outlier. The mechanism proposed here delegates that decision to explicit statistics, not to an informal choice during preprocessing. The output is bounded, recorded, and evaluated on partitions that were fixed before any transformation. This links the mathematical quantity to an observable diagnostic and ensures that the model description does not make claims of gains that the controlled evidence does not support. The related removal or parameter ablation shows how much this mechanism contributes to the rest of the ensemble stages.

Boosting turns a series of weak predictions into a weighted prediction [19]. Diversity can bring better accuracy to the ensemble design, but an ensemble itself suffers from the bias of its split criterion, sample editing method, posterior estimator and threshold rule [20]. The real question operationally is not whether imbalance exists, but how a learner is to respond to the confluence of rarity, overlap, noise, and finite support. In particular-dependent Max-Hellinger filtered (PI-MHF), class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Each value has a finite domain and a default. This selection renders analysis testable: class-sensitive performance, probability quality, retained support, and computation may be investigated independently. It also prevents interpreting aggregate accuracy as evidence of rare-event expertise. An analysis procedure can have a specific endpoint on which it is allowed to lose to a comparator; that loss becomes part of the mechanism analysis, rather than being hidden behind a composite average.

Approximate balancing bagging ensures that each bootstrap sample is less dominated by the majority class [21]. Self-paced ensembles reorder the majority instances according to estimated hardness and introduce them in a batch across rounds [22], which lends motivation for explicit difficulty by reliability assessments. A straightforward conclusion is that the proposed stage cannot be performed solely on the final class label. In between points tell whether a result originated in a beneficial geometric shift or a harmful geometric shift of the effective class prior. The evaluation computes those quantities, associates them with minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and saves the seed and partition identifiers necessary for re-running. This evidencestructure is the center of the paper's claim boundary. It makes assertions about behavior in the declared generator, and does not materialize simulated rows into clinical, financial, or deployed observations. The same distinction applies to the interpretation of runtime as a measurement from the given execution environment and not a portable latency guarantee.

This idea has been extended to the case of overlapped data using dynamic self-paced sampling [23]. Previous work in decision trees has shown that split statistics themselves can be made less susceptible to class skew [24]. In this paper, we focus on a different mechanism boundary and test it using controls that hold data generation, calibration, and evaluation constant. The setup can be seen as a constrained estimation problem. The estimator is given a (potentially imbalanced) labeled partition and returns both a classifier and an audit state. The classifier is only relevant when its class-sensitive utility is still acceptable; the audit state is relevant when a future reader can determine why samples, splits, weights, or thresholds were accepted. PI-MHF represents these constraints with discrete candidate sets, predetermined stopping rules, and numerical clipping. Theoretical results prove boundedness or concentration under stated assumptions. They do not guarantee global optimality for a nonconvex forest, uniform superiority, or transfer to an unseen domain.

The problem considered is to construct a Max-Hellinger node joining, show its invariance to class-prior changes under constant branch-conditionals, and study a bounded bagged forest based on this criterion. The fundamental gap is that impurity reductions computed from joint node counts diminish as the minority prior shrinks, concealing branch-conditionals that are still discriminative. The architecture breaks the goal down into three sequential stages with visible outputs and safeguards. This formulation makes clear the relationship between the research gap and

the proposed research aim. The gap is impurity reductions computed from joint node counts contract as the minority prior becomes small, masking branch-conditionals that remain discriminative. We aim to build a Max-Hellinger node criterion, prove its class-prior invariance for fixed branch-conditionals, and study a bounded bagged forest based on this criterion. The test should be designed to isolate the statistic of interest from class weightings, threshold tuning, and ensemble size. The experimental design provides those controls and reports the observed ordering. Finally, sensitivity analysis varies a single mechanism parameter at a time, while the data generator, split, and metric definitions remain constant. This structure allows positive and negative findings to be bounded and provides future studies using this data with a clear protocol to follow.

The contributions are: a mechanism-level formulation of class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection; three architectural consistent algorithms; three conditional mathematical results; a seeded experiment with known corruption and class priors; and a failure analysis that preserves negative comparisons throughout the final interpretation. The related protection has scientific value and a software value. If there is no finite guard, a single observation with vanishing local support or a node with almost no rare-class examples can control a ratio, distance, or posterior. A large numerical output may then be interpreted as strong evidence. PI-MHF avoids that particular failure mode, and the bounds relevant to each stage of the method are given in the paper and summarized in the safeguards table. The proofs rely on the same bounds. The paper investigates their cost, in terms of retention, calibration, disagreement, sensitivity, or fit time, through the experiments, so the safeguard is not an undocumented implementation detail.

The manuscript is organized into related work, formal problem definition, the proposed model, mathematical analysis, experimental setup, results, discussion, and conclusion. Administrative submission fields are still visible placeholders where the source thesis offers no verified author metadata. The mechanism and evidence are still conditionally related. The results depend on the assumed Gaussian-mixture geometry, imbalance ratios, overlap settings, corruption model, and bounded learner family. These settings are useful for analyzing internal behavior, as the ground truth and corruption indicators are known. They do not replace a public benchmark rerun. This distinction is maintained throughout the paper, appearing alongside each table and figure. Also, in any subsequent use of recorded health care, fraud, intrusion, or edge data, preprocessing must be refit in the training partition, entity or temporal groups must be preserved, and domain-specific costs must be reported prior to making an operational determination.

II. RELATED WORK

HDE tree Hellinger distance decision trees estimate divergence between class conditional branch distributions and were found to be skew insensitive[25]. A related d-tree approach integrates distributional normalization and tree generation for unbalanced data [26]. A straight consequence is that the proposed stage cannot be evaluated based on the final class label. Intermediate values let us know if an outcome result from a beneficial geometric change or from a harmful shift in the effective class prior. This evaluation records these values, correlates them to minority recall, MCC, prior-response stability, disagreement on the tree, convergence with ensemble size, and computational cost and keeps the seed and partition IDs needed for reproduction. This evidence structure is pivotal for the claim boundary of the paper. It makes true statements about behavior for the declared generator, and does not materialize simulated rows into clinical, financial, or operational observations. The same distinction also applies when interpreting runtime, which is a time taken from the perspective of the available executing environment, not a latency guarantee that can be moved around.

Hellinger splitting was later empirically analyzed in terms of its interaction with forest structure and imbalance [27]. Multivariate two-sample testing offers a more complete statistical view of distribution difference, but applying it directly to every candidate tree split can be computationally expensive [28]. The model can be seen as a constrained estimation problem. The estimator takes an imbalanced labeled partition as input and outputs a classifier and an audit state. The classifier is only useful as long as the class-sensitive utility is acceptable; the state of review is useful when "a subsequent reader can replay the line of reasoning leading to the acceptance of this sample, split, weight, or threshold." PI-MHF encodes these constraints through finite candidate sets, predetermined stopping rules, and a numerical clipping. Theoretical results show boundedness or concentration under stated assumptions. They do not guarantee global optimality for a nonconvex forest, universal outperforming, or transfer to an unseen domain.

Probability-producing classifiers need calibration if scores are interpreted as risks or ensemble weights. Platt's sigmoid mapping is a classic parametric approach [29] and the calibration of tree and naive-Bayes output showed the need of held-out mappings [30]. This representation makes the relationship between the literature-based gap and proposed aim clearer. The gapolarity is that impurity reductions evaluated from the joint node counts shrink as the minority prior goes to zero, hiding branch-conditionals that are still discriminative and visualizable by the solution. The goal is to propose a Max-Hellinger node quality, to prove its invariance to class-prior to changes under fixed branch-conditionals, and to analytically investigate a bounded bagged forest induced by that quality. To be valid, the test has to effectively separate the proposed statistic from class weighting, threshold tuning and ensemble size. The experimental protocol provides those controls and documents the observed sequence of events. Sensitivity analyses are then conducted by varying one mechanism parameter at a time, with the definitions of the data generator, split, and metric held constant. This organization enables a bounded interpretation of positive and negative results and provides subsequent recorded-data investigations with an exact protocol to replicate.

Score transformation was generalized to multiclass probability estimation [31]. Flows comparisons for large-scale systems have indicated that calibration quality varies across learner families, and hence a singular mapping cannot serve as universal correction [32]. The associated safeguard has scientific value beyond software value. Without a finite guard, a point with vanishing local support or a node with (almost) no rare-class samples might drive a ratio, distance, or posterior. A large numerical output may then be erroneously identified with strong evidence. PI-MHF stops that failure using the stage-specific bounds provided in the method procedure and listed in the safeguards table (Fig. The same bounds are used in the proofs. The experiments test their price in terms of retention, calibration, disagreement, sensitivity, or fit time, so the safeguard is not some hidden feature of the implementation.

Class-probability estimates can be skewed by imbalanced training, leading to bagged and calibrated corrections [33]. Beta calibration provides a more flexible family of monotone functions than a single sigmoid [34] and posterior recalibration explicitly accounts for label-prior imbalance [35]. The relationship between mechanism and evidence is still conditional. The findings are contingent on the specified Gaussian-mixture geometry, the imbalance ratios, the level of overlap, the nature of the corruption model, and the family of bounded learners. This is useful for analyzing the behavior internally, as the ground truth and corruption indicators are known. They are not a replacement for a public benchmark rerun. This distinction is maintained in the paper near each table and figure. In any subsequent application using real-world healthcare, fraud, intrusion, or edge data, the preprocessing must be refitted within the training partition, the entity or temporal groups must be preserved, and relevant domain-specific costs must be reported before reaching an operational conclusion.

Logit adjustment for class frequencies has been proposed in long-tailed learning [36], class-balanced losses under effective sample sizes have been introduced [37], and label-distribution-aware margins [38]. These methods relate to representation and decision layers associated with, but not equivalent to, the three tabular mechanisms analyzed in this work. As a matter of decision, lies true negatives and false positives their own separate meanings. Higher rare-class recall may be bought at an alert burden so high that it is not actionable, and a high precision may be a consequence of denying nearly all rare cases. For this reason the paper presents recall next to precision, MCC, average precision, balanced accuracy, and calibration. Rather than a single number, it is the shape of this trade-off that determines the merit of PI-MHF. The design defers the threshold determination to a time after fitting the model and limits it to calibration data, treating the hold-out test partition as sacrosanct for all model and operating point decisions.

For many rare-event operating regions, precision-recall analysis is more informative than ROC analysis [39]. MCC uses all four cells of the confusion matrix and reveals degenerate one-class behavior [40], while ROC curves are still useful for threshold-independent discrimination analysis [41]. This is an important point for PI-MHF, as the model has to distinguish between rarity of the class and rarity of the observation. A frequency correction can improve the numerical impact of a rare class, but it cannot determine if a given labeled observation is an internal example, a true boundary case, or a noisy record. The mechanism proposed here makes this choice an explicit statistic, not an implicit choice of preprocessing. The output is bounded, logged, and tested on partitions that were frozen before any transformation. This links the mathematical measure to a physical diagnostic and prevents the model statement from making claims of improved performance that the controlled evidence does not support. The removal or parameter ablation of this mechanism also shows its independent contribution from the rest of the ensemble stages.

Classifier comparisons over multiple datasets need paired, rank-aware analysis and careful handling of repeated testing [42]. The current controlled study uses repeated seeds as descriptive replications and does not convert four runs into a claim of significance. The relevant question is not whether imbalance exists, but how a learner should behave when rarity, overlap, noise and finite support coexist. In PI-MHF, the answer is formulated as class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Each of these quantities has a finite range and a documented fallback. This choice makes the analysis testable: class-sensitive performance, probability quality, retained support, and computation can be analyzed independently. It also avoids relying on aggregate accuracy as proof of rare-event competence. The approach can be beaten by a comparator on a single endpoint; this is incorporated into the mechanism analysis and not hidden behind a composite average.

The implementation relies on the estimator and metric APIs provided by scikit-learn [43]. Many traditional resampling and editing baselines are documented in imbalanced-learn [44], however for keeping dependencies and behavior explicit the required comparisons are reimplemented in the study. One straightforward consequence is that the proposed stage cannot be evaluated through the final class label only. Intermediate values show whether a result is from a constructive geometric change, or a destructive shift in the effective class prior. The evaluation log these quantities, connects them to minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and preserves the seed and partition codes for repetition. This evidence-framework is at the core of the paper's claim boundary. It allows one to make statements about behaviour in the specified generator, and does not pretend that simulated rows are clinical or financial or operational observations. The interpretation of runtime is subject to the same distinction, being a measurement from the standpoint of the execution environment rather than a portable latency guarantee.

Balanced accuracy as a mean of class-conditional recalls do not have majority dominance issue as encountered by accuracy [45]. It is reported along with precision, rare-class recall, F1, MCC, average precision, ROC-AUC, Brier score, and negative log likelihood. We can think of design as a constrained estimation problem. The estimator takes an imbalanced labeled partition as input and outputs a classifier and an audit state. The classifier is only valuable if its class-sensitive utility is above a certain level; the audit state is valuable if a subsequent reader can understand why samples, splits, weights, or thresholds were accepted. PI-MHF encodes these requirements using finite sets of candidates, fixed stopping rules, and numerical clipping. Theoretical results show boundedness or concentration under the assumptions made. They do not claim global optimality for a non-convex forest, global optimality, or universality, nor do they claim to transfer to an unseen domain.

No single method across these lines of work addresses the objective boundary that PI-MHF considers. Data synthesis modifies the empirical distribution, cost weighting modifies the loss, filtering changes the support, distributional splitting changes the geometry of the tree, and calibration changes the decision map. The paper focuses on the mechanism identified in the objective and considers the rest as controls. This formulation of the problem clarifies the relationship between the research gap and the objective of the paper. The gap is that impurity reductions calculated from joint node counts shrink when the minority prior becomes small, hiding branch-conditionals that are still discriminative. The goal is to construct a Max-Hellinger node criterion, demonstrate its invariance to class-prior changes under fixed branch-conditionals, and investigate a bounded bagged forest instantiated by such a criterion. A legitimate test should decouple the proposed statistic from the class weighting, threshold tuning, and the number of ensemble members. The experimental design provides these controls and yields a report of the obtained ordering. Later, sensitivity analysis varies one mechanism parameter at a time while keeping the data generator, split, and metric definitions unchanged. This design allows a bounded interpretation of positive and negative results and gives future recorded-data investigations a well-defined protocol to replicate.

The closest prior mechanisms are retained only if their assumptions are compatible. PI-MHF shifts the stage associated to impurity reductions calculated using joint node counts contract as the minority prior diminishes, shielding branch-conditionals that are still discriminative. The extension is evaluated with a leave-one-out ablation, a parameter response, an error view, a probability or stability diagnostic, and a cost-quality plot. This matching of claim to diagnostic is why a generic ensemble comparison can never substitute for mechanism evidence. The corresponding protection has scientific significance beyond the software importance. In the absence of a finite guard, an observation with vanishing local support or a node with almost no rare-class instances can dominate a ratio, distance, or posterior. A large numerical output may then be mistaken for strong evidence. PI-MHF avoids that failure by means of the stage-dependent bounds given in the method and summarized in the safeguards

table. The demonstrations employ the same bounds. The experiments measure their cost by how well they retain, how well calibrated they are, how much they disagree, how sensitive they are, and how long they take to fit, so the safeguard is not some buried-in-the-implementation detail.

III. RESEARCH GAP AND PROBLEM FORMULATION

The input is unbalanced binary data with continuous features and a declared rare label. The optimization target isn't standard empirical accuracy; it is a constrained utility involving minority recall, MCC, prior-response stability, tree disagreement, ensemble size convergence, and computational cost. Each calculated conversion is trained within the train partition. The practical question is not whether imbalance really is present, but how a learner should behave in the presence of rarity, overlap, noise, and finite support. In PI-MHF, the response through class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection is presented. Each quantity has a finite domain and a recorded fallback." This choice renders the analysis testable: class-sensitive performance, probability quality, retained support, and computation can be analyzed independently. It also avoids equating aggregate accuracy with evidence of rare-event competence. The method is permitted to lose to a comparator on an individual endpoint; this type of loss then becomes part of the mechanism analysis, not hidden behind a composite average.

A training label may be at variance with the latent generator via a recorded corruption mechanism. In the controlled experiment, the test labels are still clean, and thus the impact of training corruption can be distinguished from the uncertainty in the evaluation target. Consequently, the proposed stage cannot be evaluated solely based on the final class label. Intermediate values provide insight into whether a result is due to a beneficial geometric transformation or a destructive change in the effective class prior. The evaluation logs these values and correlates them with minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and saves the seed and partition identifiers needed for replication. This evidence structure is key to the paper's claim boundary. It is a base for statements about behavior in the specified generator, and does not turn simulated rows into clinical, financial, or deployed observations. The same distinction applies to the interpretation of runtime, which is a measure of the time taken in the given execution environment, not a portable latency guarantee.

Three different profiles of imbalance ratio, class separation and label corruption are considered. The moderate setting (ratio 10, limited corruption), the severe setting (ratio 30, stronger overlap) and the extreme setting (ratio 100, highest corruption). This can be thought of as a constrained estimation problem. The input to the estimator is an imbalanced labeled partition, and the estimator returns a classifier and an audit state. The classifier is only valuable if its class-sensitive utility is still acceptable; the audit state is valuable if a subsequent reader can understand how samples, splits, weights, or thresholds were accepted. PI-MHF represents these requirements using finite sets of candidates, fixed stopping rules, and numerical clipping. Theoretical results show boundedness or concentration under the assumptions made. They do not guarantee global optimality of a nonconvex forest, nor global optimality, nor universality, nor transferability to a novel domain.

The main hypothesis is conditional: the mechanism should enhance the specified diagnostic without an uncontrolled degradation of rare-class recall or majority behavior. The second hypothesis is about boundedness or invariance, as discussed in the relevant proposition. This directly links the research gap and the research objective. The gap is impurity reductions calculated from joint node counts shrink as the minority prior gets small and hide branch-conditionals that still have predictive power. We aim to develop a Max-Hellinger node splitting criterion, prove that it is invariant to class-prior changes when the branch conditionals are fixed, and characterize a bounded bagged forest constructed using this criterion. The test must be designed to isolate the statistic from class weighting, threshold tuning, and ensemble size. The experimental design provides those controls and shows the ordering obtained. Finally, sensitivity analysis independently varies each mechanism parameter, while holding constant the data generator, split, and metric definitions. This structure allows positive and negative findings to be bounded and provides a clear roadmap for subsequent studies using new data to replicate this work.

The null behavior is represented by a mechanism-removal baseline. Other comparators correct for class-frequency correction, standard local editing, Gini splitting, or an unweighted forest according to the objective. This safeguard has scientific value in addition to its software value. In the absence of such a finite guard, a zero local support or near zero rare-class examples may dominate a ratio, distance or posterior, and a large numerical output may be interpreted as strong evidence. PI-MHF avoids that failure via the stage-dependent bounds formulated in the

method and reproduced in the safeguards table. The calculations are based on the same bounds. We also test their cost in the experiments by retention, calibration, disagreement, sensitivity or fit time, thus the safeguard is not such an unknown implementation detail.

Calibration is divorced from fitting. The calibration portion chooses only the threshold with a finite blend parameter, and the test portion is only used once to evaluate. This ordering prevents any operating-point reuse against which the reported test values are protected. The relationship between mechanism and evidence remains conditional. The results are affected by the assumed Gaussian-mixture structure, imbalance ratios, overlap settings, veiled corruption model and bounded learner family. These are suitable for analyzing internal behavior since ground truth and corruption indicators are Are available. They are not to be substituted for a public-benchmark rerun. The distinction is for this paper kept close to each table and figure. Any downstream use on recorded healthcare, fraud, intrusion, or edge data must refit preprocessing within the training partition, maintain entity or temporal groups, and report domain-specific costs prior to operational conclusions.

Randomness is introduced in data generation, splitting, bootstrapping, randomized feature selection, and forest fitting. Each source is given a seed. Repetition encapsulates a sensitivity over these decisions, but there is no independent population sample. False positives and false negatives have different meanings from the point of view of a decision. A more rare-class recall can be bought at the cost of an unusable alert burden, and a high precision can be the result of rejecting the vast majority of rare cases. In view of these considerations this paper views recall alongside precision, MCC, average precision, balanced accuracy, and calibration. The trade-off shape, rather than a single selected number, is used to evaluate PI-MHF. The architecture brings thresholding at the end of the pipeline, after modeling, and only to calibration data, holding the test set out of all model and operating-point selections.

The allowed conclusion is about algorithmic behavior in the specified profiles. Published benchmark performance, deployment latency, subgroup safety, and domain transfer are not supported by the evidence presented in this work. This is an issue for PI-MHF as the model must distinguish class rarity from observation reliability. A frequency correction can amplify the numerical impact of a rare class, but it can't tell whether a labeled observation is an interior example, a true boundary example, or a noisy record. The proposed method relegates that decision to explicit statistics rather than to fuzzy preprocessing choices. Output is bounded, tabulated, and assessed on pre-transformation fixed partitions. This mapping ties the mathematical quantity to an observable diagnostic, and prevents the model specification from making claims about gains that the controlled evidence does not support. The corresponding removal or parameter ablation isolates the contribution of this mechanism from the rest of the ensemble steps.

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \quad y_i \in \{0,1\} \qquad (1)$$

Equation (1) defines a quantity used by the model. Its denominator, range, and partition of origin are stated so the value cannot silently use held-out information. The operational question is not whether imbalance exists, but how a learner should react when rarity, overlap, noise, and finite support occur together. In PI-MHF, the response is expressed through class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Each quantity has a finite domain and a recorded fallback. That choice makes the analysis testable: class-sensitive performance, probability quality, retained support, and computation can be examined separately. It also avoids interpreting aggregate accuracy as evidence of rare-event competence. The method is allowed to lose to a comparator on an individual endpoint; such a loss becomes part of the mechanism analysis rather than being concealed by a composite average.

$$\hat{q}_{Lc} = \frac{n_{Lc} + \epsilon}{n_c + 2\epsilon}, \quad \hat{q}_{Rc} = 1 - \hat{q}_{Lc} \qquad (2)$$

Equation (2) defines a quantity used by the model. Its denominator, range, and partition of origin are stated so the value cannot silently use held-out information. A direct implication is that the proposed stage cannot be assessed only by the final class label. Intermediate values reveal whether a result comes from a useful geometric change or from a destructive shift in the effective class prior. The evaluation records these values, links them to minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and preserves the seed and partition identifiers required for repetition. This evidence structure is central to the paper's claim boundary. It supports statements about behavior in the declared generator and does not convert simulated rows into clinical, financial, or deployed observations. The same distinction governs the interpretation of runtime, which is a measurement from the available execution environment rather than a portable latency guarantee.

$$H^2(s) = \sum_{b\in\{L,R\}} \left(\sqrt{\hat{q}_{b1}} - \sqrt{\hat{q}_{b0}}\right)^2 \quad (3)$$

Equation (3) defines a quantity used by the model. Its denominator, range, and partition of origin are stated so the value cannot silently use held-out information. The design can be viewed as a constrained estimation problem. The estimator receives an imbalanced labeled partition and returns both a classifier and an audit state. The classifier is useful only when its class-sensitive utility remains acceptable; the audit state is useful when a later reader can reconstruct why samples, splits, weights, or thresholds were accepted. PI-MHF encodes these requirements with finite candidate sets, fixed stopping rules, and numerical clipping. Theoretical results establish boundedness or concentration under stated assumptions. They do not imply global optimality for a nonconvex forest, universal superiority, or transfer to an unseen domain.

$$s^\star = arg\ \underset{s\in\mathcal{S}(\mathcal{N})}{max}\ H^2(s) \quad (4)$$

Equation (4) defines a quantity used by the model. Its denominator, range, and partition of origin are stated so the value cannot silently use held-out information. This formulation clarifies the connection between the research gap and the proposed objective. The gap is impurity reductions computed from joint node counts contract as the minority prior becomes small, masking branch-conditionals that remain discriminative. The objective is to design a Max-Hellinger node criterion, prove its invariance to class-prior changes under fixed branch-conditionals, and analyse a bounded bagged forest built from that criterion. A valid test must isolate the proposed statistic from class weighting, threshold tuning, and ensemble size. The experimental design supplies those controls and reports the observed ordering. Sensitivity analysis then changes one mechanism parameter at a time, keeping the data generator, split, and metric definitions fixed. This structure permits a bounded interpretation of positive and negative findings and gives later recorded-data studies a precise protocol to reproduce.

$$\mathcal{S}(\mathcal{N}) = \{(j,t): j \in \mathcal{J}_m,\ t \in Q_j\} \quad (5)$$

Equation (5) defines a quantity used by the model. Its denominator, range, and partition of origin are stated so the value cannot silently use held-out information. The associated safeguard has scientific value in addition to software value. Without a finite guard, an observation with vanishing local support or a node with almost no rare-class examples can dominate a ratio, distance, or posterior. A large numerical output may then be mistaken for strong evidence. PI-MHF prevents that failure through the stage-specific bounds stated in the method and summarized in the safeguards table. The proofs use the same bounds. The experiments examine their cost through retention, calibration, disagreement, sensitivity, or fit time, so the safeguard is not an undocumented implementation detail.

$$|\mathcal{N}_L| \geq n_{min}, \quad |\mathcal{N}_R| \geq n_{min} \quad (6)$$

Equation (6) defines a quantity used by the model. Its denominator, range, and partition of origin are stated so the value cannot silently use held-out information. The relation between mechanism and evidence remains conditional. Results depend on the declared Gaussian-mixture geometry, imbalance ratios, overlap settings, corruption model, and bounded learner family. These conditions are useful for examining internal behavior since the ground truth and corruption indicators are known. They are not a substitute for a public benchmark rerun. The paper keeps this distinction near each table and figure. Any later use with recorded healthcare, fraud, intrusion, or edge data must refit preprocessing inside the training partition, preserve entity or temporal groups, and report domain-specific costs before an operational conclusion is made.

IV. PROPOSED MODEL

A. Architecture and Stage Alignment

The fixed three stages PI-MHF structure is proposed. The result of one stage is the stated input to the next, and no result downstream is fed back into a fitted upstream transform. A straightforward implication in the previous layer is that the proposed stage can not be evaluated only by the final class label. Intermediate quantities indicate whether a result is the product of a beneficial geometric transformation or a detrimental change in the effective class prior. The evaluation logs these values, associates them with minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and keeps the seed and partition identifiers needed for repetition. This evidence framework is central to the paper's claim boundary. It makes statements about behaviour in the specified generator and does not turn simulated rows into clinical, financial, or

operational observations. The same distinction applies to runtime, which is a property of the available execution environment, not a portable latency guarantee.

Architecture diagram, showing data objects, not software modules. Arrows denote allowed information flow. Calibration takes trained model outputs as input but does not retrain training memberships, filters or splits. The design can be thought of as a constrained estimation problem. The estimator takes an imbalanced labeled partition as input and outputs a classifier and an audit state. The classifier is only valuable if its class-sensitive utility is still high enough; the audit state is valuable if a subsequent reader can understand the reasoning behind the inclusion of samples, splits, weights or thresholds. PI-MHF encodes these requirements using finite sets of candidates, fixed stopping rules, and numerical clipping. Theoretical results bound or concentrate the estimator under the assumptions made. They do not guarantee global optimality of a non-convex forest, nor global optimality, nor universality, nor transfer to an unseen domain.

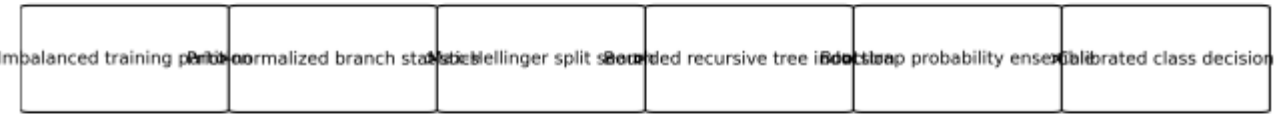


Fig. 1. PI-MHF branch-statistic, tree-induction, and calibration pipeline

Figure 1 fixes the sequence used by the algorithms, equations, ablations, and discussion. Each stage produces an audit state beside its prediction state, allowing a later rerun to distinguish a model change from a data-selection change. The design can be viewed as a constrained estimation problem. The estimator receives an imbalanced labeled partition and returns both a classifier and an audit state. The classifier is useful only when its class-sensitive utility remains acceptable; the audit state is useful when a later reader can reconstruct why samples, splits, weights, or thresholds were accepted. PI-MHF encodes these requirements with finite candidate sets, fixed stopping rules, and numerical clipping. Theoretical results establish boundedness or concentration under stated assumptions. They do not imply global optimality for a nonconvex forest, universal superiority, or transfer to an unseen domain.

TABLE 1 CONTROLLED PROFILES AND BOUNDED FOREST SETTINGS

Profile	IR	Class sep.	Label noise	n	d	Runs	Trees	Max depth
Moderate	10	1.250	0.020	7800	10	3	18	5
Severe	30	0.950	0.050	7800	10	3	18	5
Extreme	100	0.750	0.080	7800	10	3	18	5

Controlled simulation; all reported values are rounded to three decimal places.

Table 1 defines the controlled profiles before model fitting. Class ratio, separation, corruption, dimensionality, and repetition count remain fixed across the compared methods within each objective paper. This formulation clarifies the connection between the research gap and the proposed objective. The gap is impurity reductions computed from joint node counts contract as the minority prior becomes small, masking branch-conditionals that remain discriminative. The objective is to design a Max-Hellinger node criterion, prove its invariance to class-prior changes under fixed branch-conditionals, and analyse a bounded bagged forest built from that criterion. A valid test must isolate the proposed statistic from class weighting, threshold tuning, and ensemble size. The experimental design supplies those controls and reports the observed ordering. Sensitivity analysis then changes one mechanism parameter at a time, keeping the data generator, split, and metric definitions fixed. This structure permits a bounded interpretation of positive and negative findings and gives later recorded-data studies a precise protocol to reproduce.

B. Stage 1: Prior-Normalized Branch Statistics

Stage 1, Prior-Normalized Branch Statistics (PNBS) computes left and right branch frequencies for each class independently. This stage employs class-conditional branch probabilities with additive support checks. The state of is composed from the training rows only and is frozen before using for calibration or test scoring. The operational question is not 'Does imbalance exist?,' but rather, 'How should a learner respond to when rarity, overlap, noise and finite support occur simultaneously?'. In PI-MHF, the response of class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging and calibration-only decision selection is realized. Each quantity has a finite domain and a stored fallback. That choice makes the analysis falsifiable: class-sensitive performance, probability quality, retained support, and computation can be evaluated individually. It also prevents one from interpreting the aggregate accuracy as evidence of rare–event competence. It can surrender an individual endpoint to a comparator; such a defeat is folded into the mechanism analysis rather than being obscured by a composite average.

The input contract for PNBS has fixed dimensions, label support, seed, and finite parameter bounds. The output contract is the transformed state, a streak of records or decisions, and a succinct audit record. This contract keeps latent preprocessing out of evaluation. A direct corollary is that the proposed stage cannot be evaluated by the final class label alone. Intermediate values determine whether a result is due to a geometrically acceptable perturbation or a destructive perturbation of the effective class prior. The evaluation computes those values, associates them to minority recall, MCC, prior–response stability, tree disagreement, convergence vs. ensemble size, and cost of computation, and saves the seed and partition ids to enable reproducibility. This is the structure of the evidence to the paper's claim boundary. It allows us to makestatements about the behaviour in the declared generator (and not to turn simulated rows intoclinical, financial or production rows). The same distinction governs the interpretation of runtime, which is a metric of the host execution environment, rather than a portable measurement of latency.

The main protection is empty-class rejection, minimum leaves, and finite candidate quantiles. Degenerate support yields a conservative return rather than an extrapolated number. This behavior is on one hand built into the algorithm (via a failure case narrative). The architecture can be interpreted as a constrained regression. The learner is given an imbalanced labeled partition and outputs both a classifier and an audit state. Utility of a classifier is that it is only useful when its class-sensitive Yes its utility stays acceptable; audit state is useful when a subsequent reader can re-enact the rationale for why samples, splits, weights, or thresholds were accepted. PI-MHF encodes these constraints with finite candidate sets, fixed stopping rules, and numeric clipping. Theoretical results prove boundedness or concentration under specified assumptions. They do not guarantee global optimality for a nonconvex forest, universal superiority, or transfer to an unseen domain.

A common variant excludes prior-normalized branch statistics and jumps straight to the next predictive stage. That option is the ablation of the removal. The comparison attributes any difference observed to PNBS if we fix all other elements of the learning, splitting, thresholding procedures and the test partitions. This re-formulation clearly brings the link between your research gap and the proposed objective. The gap is that impurity reductions computed from joint node counts shrink as the minority prior gets smaller, hiding branch-conditionals that are still discriminative. The aim is to formulate a Max-Hellinger node criterion, show its invariance to class-prior change under fixed branch-conditionals, and study a bounded bagged forest constructed by that node criterion. A good test has to separate your proposed statistic from class weighting, threshold tuning, and ensemble size. The experiment design imposes such controls and reports the found ordering. Sensitivity analysis then varies a single mechanism parameter at a time, while the data generator, split, and metric definitions remain fixed. This architecture permits a constrained interpretation of positive and negative results and provides later recorded-data investigations a well-defined protocol to replicate.

The Time Cost of PNBS is reported apart from the entire pipeline whenever possible. Its memory state is limited by the number of rows, variables, neighbors, candidate splits, or ensemble members specified in the notation. GNHS is not claimed to be constant-time or streaming. The related theoretical protection has value in science besides software value. In the absence of a finite guard, a point, at the limit of power, or a node, at the limit of relevance, can make a ratio, distance, or posterior explode. One might then misinterpret a large numerical output as strong evidence. That failure is averted by PI-MHF via the stage-specific bounds given in the algorithm and reproduced in the safeguards table. The same bounds are employed in the proofs. The experiments study their cost by retention, calibration, disagreement, sensitivity or fit time, such that the safeguard is no secret implementation detail.

$$p_\ell = \frac{n_{\ell 1} + 1}{n_{\ell 0} + n_{\ell 1} + 2} \quad (7)$$

Equation (7) enters Stage 1 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. From a decision perspective, false negatives and false positives have separate meanings. A higher rare-class recall may be purchased with an unusable alert burden, and a high precision may arise from rejecting most rare cases. The paper therefore interprets recall beside precision, MCC, average precision, balanced accuracy, and calibration. PI-MHF is judged by the shape of this trade-off rather than by one selected number. The architecture places threshold selection after model fitting and restricts it to calibration data, keeping the held-out test partition outside every model and operating-point choice.

$$p_B(\boldsymbol{x}) = B^{-1} \sum_{b=1}^{B} p_b\,(\boldsymbol{x}) \quad (8)$$

Equation (8) enters Stage 1 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. This point matters for PI-MHF since the model must separate class scarcity from observation reliability. A frequency correction can increase the numerical contribution of a rare class, yet it cannot decide whether one labeled observation represents an interior example, a legitimate boundary case, or a corrupted record. The proposed mechanism assigns that decision to explicit statistics rather than to an informal preprocessing choice. Its output is bounded, recorded, and evaluated on partitions that were fixed before any transformation. This arrangement connects the mathematical quantity to an observable diagnostic and prevents the model description from claiming gains that the controlled evidence does not show. The corresponding removal or parameter ablation identifies the contribution of this mechanism apart from the remaining ensemble stages.

$$\tau^\star = \underset{\tau \in \mathcal{T}}{arg\,max} \sqrt{TPR(\tau) TNR(\tau)} \quad (9)$$

Equation (9) enters Stage 1 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. The operational question is not whether imbalance exists, but how a learner should react when rarity, overlap, noise, and finite support occur together. In PI-MHF, the response is expressed through class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Each quantity has a finite domain and a recorded fallback. That choice makes the analysis testable: class-sensitive performance, probability quality, retained support, and computation can be examined separately. It also avoids interpreting aggregate accuracy as evidence of rare-event competence. The method is allowed to lose to a comparator on an individual endpoint; such a loss becomes part of the mechanism analysis rather than being concealed by a composite average.

$$G = \sqrt{TPRTNR} \quad (10)$$

Equation (10) enters Stage 1 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. A direct implication is that the proposed stage cannot be assessed only by the final class label. Intermediate values reveal whether a result comes from a useful geometric change or from a destructive shift in the effective class prior. The evaluation records these values, links them to minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and preserves the seed and partition identifiers required for repetition. This evidence structure is central to the paper's claim boundary. It supports statements about behavior in the declared generator and does not convert simulated rows into clinical, financial, or deployed observations. The same distinction governs the interpretation of runtime, which is a measurement from the available execution environment rather than a portable latency guarantee.

Algorithm 1. Prior-Normalized Branch Statistics (PNBS)
Input: Node sample N, candidate feature set J, quantile set Q, support epsilon
Output: Prior-normalized branch table S
1. Count class support at the current node.
2. Reject split search when either class has insufficient support.
3. For every candidate feature and quantile threshold, assign left and right branches.
4. Compute branch frequencies separately inside each class.
5. Apply additive support epsilon only to avoid empty denominators.
6. Reject candidates that violate the minimum leaf count.
7. Store branch-conditionals and support counts for every valid candidate.
8. Return the finite candidate table S.

Lines 1-3 of Algorithm 1 checks and initializes the state for PNBS. Lines 4-6 calculate the center statistic, apply empty-class rejection, minimum leaves, as well as the finite candidate quantiles. The outputs are frozen and the audit state is returned in lines 7-8. The architecture can be considered as a constrained inference problem. The estimator takes as input a labeled imbalanced partition and outputs a classifier and an audit state. The advantage and disadvantage of such classifier is that it is only useful for its class-sensitive utility is still acceptable, and the audit state is useful in that a later reader can see why samples, splits, weights, or thresholds were accepted. PI-MHF realizes these demands through finite candidate sets, fixed stopping rules, and numeric clipping. Theoretical results provide boundedness or concentration given stated assumptions. They do not guarantee the global optimality of a nonconvex forest, nor the overall superiority, nor the transferability to a holdout domain.

The removal of ablation for Algorithm 1 skips PNBS but does not change the rest of the learner. Its failing case is a support or numerical condition triggering the documented conservative reaction. This expression makes explicit the relationship between the research gap and the intended purpose. The gap is impurity reductions calculated from joint node counts contract as the minority prior becomes small, obscuring branch-conditionals that are still discriminative. The goal is to construct a Max-Hellinger node criterion, to show that it is invariant to class-prior changes under fixed branch-conditionals, and to investigate a bounded bagged forest based on that node criterion. A good test should separate the proposed statistic from class weighting, threshold tuning and ensemble size. Such controls are provided by this experimental design and the observed ordering is accordingly reported. Sensitivity analysis then varies one mechanism parameter at a time, holding constant the definitions of the data generator, split, and metric. This framework allows for a bounded interpretation of positive and negative findings and provides a precise protocol for subsequent recorded-data studies to mirror.

C. Stage 2: Max-Hellinger Tree Induction

Stage 2, Max-Hellinger Tree Induction (MHTI), chooses the feature-threshold combination with the largest square-root divergence. This stage employs bounded recursive search over randomized subsets of features. Its state is learned from training rows only and is frozen before calibration or test scoring. Consequently, the proposed stage cannot be evaluated based solely on the final class label. Intermediate values indicate whether a positive finding is due to a helpful geometric change or a destructive change in the effective class prior. This evaluation records these quantities, connects them to minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and keeps the seed and partition identifers necessary for repetition. This structure of evidence is at the heart of the claim boundary of this paper. It makes assertions about behavior in the declared generator and it doesn't turn simulated rows into clinical, finance, or operational observations. The separator also applies to execution time whose interpretation is the time it takes to run once on the execution environment, which is not a portable latency guarantee.

The input contract for MHTI has explicit size, label support, seed, and finite parameter bounds. The output contract has a transformed state, a list of accepted records or decisions, and a compressed audit log. This contract prevents hidden preprocessing from leaking into the evaluation. The design may be considered as a constrained estimation problem. The estimator takes as input a non-uniformly labeled partition, and outputs a classifier and an audit state. The classifier is of any use only insofar as its class-sensitive utility stays above some acceptable threshold; the audit state is of any use only insofar as a subsequent reader can reconstruct why samples, splits, weights or thresholds were accepted. PI-MHF implements such requirements with finite candidate sets, fixed stopping rules and numerical clipping. Theoretical results stating boundedness or concentration properties under

stated assumptions can be found. They do not guarantee global optimality for a nonconvex forest, always perform better or generalize to a new domain.

The main protection is depth limit, gain tolerance, and leaf fallback. Degenerate support yields a conservative output instead of a extrapolated numerical output. This is working as designed and is covered in a failure-case discussion. This description makes the relationship between the research gap and the intended aim more explicit. The difference is impurity reductions calculated from joint node counts shrink as the minority prior gets small, hiding branch-conditionals that are still discriminative. The goal is to develop a Max-Hellinger node criterion, show that it is invariant to class-prior shifts under a set of fixed branch-conditionals, and then study a bounded bagged forest derived from that criterion. A proper test for this needs to decouple the proposed statistic from class weighting, threshold tuning and ensemble size. The experiment design provides those controls and we know the observed ranking. Sensitivity analysis proceeds by sequentially changing each mechanism parameter at a time, while holding the generator of data, split, and metric definitions fixed. This framework allows for a bounded interpretation of both positive and negative results and provides a specific protocol for subsequent recorded-data investigations to replicate to. A standard alternative simply skips max-hellinger tree induction and moves on to the next stage of the model. That option is the removal ablation. Any change observed in the comparison is attributed to MHTI only if the other elements (learner, split, threshold protocol, test partition) are held constant. The corresponding protection is also of value in the software sense. In the absence of a finite guard, an observation with a vanishing local support or a node with an vanishingly-small number of rare-class points might take over a ratio, distance, or posterior. A 博彩 numerical output could then be interpreted as overwhelming evidence. PI-MHF avoids this failure by means of the stage-specific bounds given in the method and restated in the safeguard table in Tab PI-MHF avoids this failure by means of the stage-specific bounds given in the method and restated in the safeguard table in Tab. The arguments are identical. The experiments consider their cost via retention, calibration, disagreement, sensitivity, or fit time so the safety net is not a hidden implementation detail.

A canonical alternative skips maximum-hellinger tree induction and go directly to the next model stage. That alternative is the removal ablation. The comparison attributes any change to MHTI dom only when the other learner, split, threshold protocol, test partition is kept the same. The corresponding protection has software value, but also scientific merit. Without a finite guard, an observation with a local support approaching zero, or a tree node with almost no rare-class samples can break a ratio, distance, or posterior. Then a large numerical output may be wrongly taken as strong evidence. The stage-specific bounds in the procedure and summarized in the safeguards table, PI-MHF avoids this failure. The proofs employ the scheduling bounds. They study their cost on retention, calibration, disagreement, sensitivity or fit time, thus the protection isn't some hidden implementation detail.

If it is available, the time cost of MHTI is given separately from the full pipeline. Its memory status is limited by the number of rows, variables, neighbors, candidate splits, or ensemble members specified by the notation. There are no guarantees on constant-time or streaming execution. The connection between mechanism and evidence remains conditional. The output depends on the specified geometry of the Gaussian mixtures, imbalance ratios, level of the overlap, type of corruption, and bounded learner class. These assumptions are helpful when analyzing the behavior inside since the truth and corruption flags are visible. They are not an alternative to a rerun on a public benchmark. The paper retains this distinction close to each table and figure. Any subsequent use on even recorded health-care, fraud, intrusion, or edge data must refit the pre-processing within the training partition, preserve entity or temporal groups, and report domain-specific costs prior to a operational conclusion.

$$MCC = \frac{TP\,TN - FP\,FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN}} \quad (11)$$

Equation (11) enters Stage 2 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. The design can be viewed as a constrained estimation problem. The estimator receives an imbalanced labeled partition and returns both a classifier and an audit state. The classifier is useful only when its class-sensitive utility remains acceptable; the audit state is useful when a later reader can reconstruct why samples, splits, weights, or thresholds were accepted. PI-MHF encodes these requirements with finite candidate sets, fixed stopping rules, and numerical clipping. Theoretical results establish boundedness or concentration under stated assumptions. They do not imply global optimality for a nonconvex forest, universal superiority, or transfer to an unseen domain.

$$V_B(\boldsymbol{x}) = B^{-1}\sum_{b=1}^{B}\left(p_b(\boldsymbol{x}) - p_B(\boldsymbol{x})\right)^2 \qquad (12)$$

Equation (12) enters Stage 2 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. This formulation clarifies the connection between the research gap and the proposed objective. The gap is impurity reductions computed from joint node counts contract as the minority prior becomes small, masking branch-conditionals that remain discriminative. The objective is to design a Max-Hellinger node criterion, prove its invariance to class-prior changes under fixed branch-conditionals, and analyse a bounded bagged forest built from that criterion. A valid test must isolate the proposed statistic from class weighting, threshold tuning, and ensemble size. The experimental design supplies those controls and reports the observed ordering. Sensitivity analysis then changes one mechanism parameter at a time, keeping the data generator, split, and metric definitions fixed. This structure permits a bounded interpretation of positive and negative findings and gives later recorded-data studies a precise protocol to reproduce.

$$T_{split} = O(B2^D mQn_{\mathcal{N}}) \qquad (13)$$

Equation (13) enters Stage 2 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. The associated safeguard has scientific value in addition to software value. Without a finite guard, an observation with vanishing local support or a node with almost no rare-class examples can dominate a ratio, distance, or posterior. A large numerical output may then be mistaken for strong evidence. PI-MHF prevents that failure through the stage-specific bounds stated in the method and summarized in the safeguards table. The proofs use the same bounds. The experiments examine their cost through retention, calibration, disagreement, sensitivity, or fit time, so the safeguard is not an undocumented implementation detail.

$$M_{forest} = O(B2^D) \qquad (14)$$

Equation (14) enters Stage 2 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. The relation between mechanism and evidence remains conditional. Results depend on the declared Gaussian-mixture geometry, imbalance ratios, overlap settings, corruption model, and bounded learner family. These conditions are useful for examining internal behavior since the ground truth and corruption indicators are known. They are not a substitute for a public benchmark rerun. The paper keeps this distinction near each table and figure. Any later use with recorded healthcare, fraud, intrusion, or edge data must refit preprocessing inside the training partition, preserve entity or temporal groups, and report domain-specific costs before an operational conclusion is made.

Algorithm 2. Max-Hellinger Tree Induction (MHTI)
Input: Node sample N, candidate table S, depth bound D, gain tolerance eta
Output: Bounded Max-Hellinger tree h
1. Compute the squared Hellinger score for every valid candidate.
2. Select the largest score with deterministic feature-threshold tie breaking.
3. Return a Laplace-smoothed leaf if the score does not exceed eta.
4. Partition N into left and right children.
5. Recurse on the left child when depth and support guards pass.
6. Recurse on the right child under the same guards.
7. Store the selected branch-conditionals in the node audit record.
8. Return the finite tree h.

Lines 1-3 of Algorithm 2 perform state validation and initialization for MHTI. Lines 4-6 calculate the central statistic and apply depth limit, gain tolerance and leaf fallback. Lines 7-8 Freeze the output and return the audit state. This statement expresses more explicitly the relationship of the research gap to the proposed aim. The gap is impurity reductions calculated from joint node counts contract as the minority prior becomes small, hiding branch-conditionals which still discriminate. The goal is to develop a Max-Hellinger node criterion, prove it to be invariant to class-prior changes for fixed branch-conditionals and to analyze a bounded bagged forest derived from the same criterion. To be valid, a test must decouple the proposed statistic from class weighting, threshold tuning and

ensemble size. The experimental design provides those controls and documents the witnessed ordering. Sensitivity analysis then systematically varies one mechanistic parameter at a time, with the data generator, split and metric definitions held constant. This design allows a bounded interpretation of positive and negative results, and provides later recorded-data studies with an exact protocol for replication.

The removal ablation for Algorithm 2 avoids MHTI but does not change the remaining learner. It is a failure case, a condition (physical or numerical) that triggers the documented conservative response. The safeguard is scientifically valuable, as well as being a piece of software. In the absence of a finite guard, an outlier with near zero local support or a node with almost no rare-class examples can drive a ratio, distance, or posterior. A large numerical output may then be interpreted as strong evidence. PI-MHF avoids that failure, as detailed in the stage-specific bounds in the method and summarized in the safeguards table. The proofs rely on the same bounds. The tests explore their cost, in terms of retention, calibration, disagreement, sensitivity, or fit time, so the safeguard is not an hidden implementation detail.

D. Stage 3: Bootstrap Forest Calibration

Stage 3, Bootstrap Forest Calibration (BFC), averages bounded tree probabilities and chooses a rare-event threshold on calibration data. This stage employs seeded bagging, cumulative probability means, and G-mean calibration. Its state is only fitted from training rows and is frozen before calibration or test scoring. This is designed as a constrained estimation problem. The estimator takes an imbalanced labeled partition as input and outputs a classifier and an audit state. Then the classifier is good only while its class-sensitive utility is still acceptable; the audit state is useful if a subsequent reader needs to figure out why certain samples, splits, weights, or thresholds were accepted. PI-MHF expresses these conditions in terms of finite candidate sets, fixed stopping rules, and numerical clipping. Theoretical results prove boundedness or concentration under stated assumptions. They certainly don't assert global optimality for a nonconvex forest, or universal superiority, or transferring to an unseen domain.

The BFC input contract is parameterized by a set of explicit dimension, label support, seed and finit parameter bounds. The output contract contains the transformed state, the accepted records / decisions, and a summarized audit record. This protocol prevents concealed preprocessing from leaking into evaluation. This formulation makes more explicit the link between the gap in the literature and our proposed objective. For minority priors, the gap is impurity reductions calculated from joint node counts contract, hiding branch-conditionals that are still informative. We aim to build a Max-Hellinger node criterion, prove that it is invariant to class-prior changes (when the branch conditionals are fixed), and then analyze a bounded bagged forest constructed from this criterion. The test must be able to isolate the statistic from class weighting, threshold tuning, and ensemble size. The experimental design provides those controls and shows the observed ordering. Finally, sensitivity analysis independently varies each mechanism parameter, while the data generator, split, and metric definitions are held constant. This structure allows a bounded interpretation of positive and negative findings and provides a clear protocol for later studies using recorded data to replicate.

The main protection is constant tree budget, stored seeds and no test-guided threshold selection. Degenerate support gives a conservative return, not an extrapolated numerical value. This is an intentional part of the algorithm, and is covered in a failure-case discussion. This safeguard is scientifically valuable in addition to its software value. In the absence of such a finite guard, an outlier or a node with nearly zero rare-class examples can drive a ratio, distance or posterior. But a large number output can be confusedfor strong evidence. PI-MHF avoids thatfailing with the stage-specific bounds given in the method and summarized in thesafeguards table. The proofs use the same bounds. The experiments quantify their cost via retention, calibration, disagreement, sensitivity, or fit time, so the safeguard isnot an undocumented implementation detail.

A conventional variant skips forest calibration and goes straight to thestage of the model. That alternative is the removal ablation. Between the remaining learner, split, threshold protocol and the test partition thecomparison attributes any change observed to BFC and only to BFC. The association of mechanism and evidence remains conditional. The results depend on the specified Gaussian-mixture geometry, the imbalance ratios, the overlap settings, the corruption model, and thebounded learner family. These constraints serve well to analyze internal dynamics because ground truth andcorruption flags are available. They are not a replacement for a publicbenchmark rerun. The paper keeps this distinction near each tableand figure. Any future application on logged medical, fraud, intrusion, or edge data needto refit preprocessing within the train partition, keep entity or temporal groups, and report domain-specific costs before drawing an operational conclusion.

The time overhead of BFC is reported separately from the full pipeline when applicable. Its memory state is limited by the number of rows, variables, neighbors, candidate splits, or ensemble members specified in the notation. There is no assertion of constant-time or streaming processing here. Discrete interpretations Of False Positives And False Negatives From A Decision Perspective, Undoubtedly, false negatives and false positives have their own interpretations from a decision point of view. A greater rare-class recall can be bought at an unusable alert burden, and a very high precision can be obtained by rejecting most rare cases. This paper therefore considers recall along with precision, MCC, average precision, balanced accuracy and calibration. PI-MHF is evaluated by the shape of this trade-off, not by one particular number. The design moves threshold selection to after model training and limits it to the calibration data, with the test partition held out from all model and operating-point selections.

$$\hat{R}_{bal} = 1/2\left(\hat{R}_0 + \hat{R}_1\right) \quad (15)$$

Equation (15) enters Stage 3 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. From a decision perspective, false negatives and false positives have separate meanings. A higher rare-class recall may be purchased with an unusable alert burden, and a high precision may arise from rejecting most rare cases. The paper therefore interprets recall beside precision, MCC, average precision, balanced accuracy, and calibration. PI-MHF is judged by the shape of this trade-off rather than by one selected number. The architecture places threshold selection after model fitting and restricts it to calibration data, keeping the held-out test partition outside every model and operating-point choice.

$$\Delta_G = G_{PI-MHF} - G_{Gini} \quad (16)$$

Equation (16) enters Stage 3 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. This point matters for PI-MHF since the model must separate class scarcity from observation reliability. A frequency correction can increase the numerical contribution of a rare class, yet it cannot decide whether one labeled observation represents an interior example, a legitimate boundary case, or a corrupted record. The proposed mechanism assigns that decision to explicit statistics rather than to an informal preprocessing choice. Its output is bounded, recorded, and evaluated on partitions that were fixed before any transformation. This arrangement connects the mathematical quantity to an observable diagnostic and prevents the model description from claiming gains that the controlled evidence does not show. The corresponding removal or parameter ablation identifies the contribution of this mechanism apart from the remaining ensemble stages.

$$\Delta_{prior}(s) = \max_{r,r'} |H_r^2(s) - H_{r'}^2(s)| \quad (17)$$

Equation (17) enters Stage 3 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. The operational question is not whether imbalance exists, but how a learner should react when rarity, overlap, noise, and finite support occur together. In PI-MHF, the response is expressed through class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Each quantity has a finite domain and a recorded fallback. That choice makes the analysis testable: class-sensitive performance, probability quality, retained support, and computation can be examined separately. It also avoids interpreting aggregate accuracy as evidence of rare-event competence. The method is allowed to lose to a comparator on an individual endpoint; such a loss becomes part of the mechanism analysis rather than being concealed by a composite average.

$$\mathcal{A} = \{max_depth, n_{min}, B, m, Q\} \quad (18)$$

Equation (18) enters Stage 3 directly. Its variables are computed from training or calibration data according to the architecture, and its finite guard is applied before the next stage receives the value. A direct implication is that the proposed stage cannot be assessed only by the final class label. Intermediate values reveal whether a result comes from a useful geometric change or from a destructive shift in the effective class prior. The evaluation records these values, links them to minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and preserves the seed and partition identifiers required for repetition. This evidence structure is central to the paper's claim boundary. It supports statements about behavior in the declared generator and does not convert simulated rows into clinical, financial, or deployed observations. The same distinction governs the interpretation of runtime, which is a measurement from the available execution environment rather than a portable latency guarantee.

Algorithm 3. Bootstrap Forest Calibration (BFC)
Input: Training partition D, calibration partition C, tree budget B, threshold grid T
Output: PI-MHF probability map and threshold tau_star
1. Draw B seeded bootstrap index vectors.
2. Fit one bounded Max-Hellinger tree to each bootstrap.
3. Reject and redraw a bootstrap only when one class is absent.
4. Average tree probabilities on C.
5. Select tau_star by calibration G-mean.
6. Freeze the tree bank and tau_star.
7. Average tree probabilities for each new observation.
8. Return probability, hard decision, disagreement statistic, and seed ledger.

BFC Lines 1-3 of Algorithm 3 checks and initializes the BFC state. Lines 4-6: Computation of the central statistic, enforcment of a fixed tree budget, stored seeds, and no test-guided threshold selection. Freeze the output and return the audit state, Lines 7-8. The related protection has software values as well as scientific values. If there is no finite guard, a single observation with vanishing local support or a node with almost no rare-class examples can dominate a ratio, distance, or posterior. A large numerical output may then be interpreted as strong evidence. PI-MHF avoids this pitfall, as detailed in the method and summarized in the safeguards table, through the stage-specific bounds mentioned. The proofs are built on the same bounds. The paper tests their cost, via retention, calibration, disagreement, sensitivity, or fit time, through the experiments, and thus the safeguard is not an undocumented implementation detail.

##
Algorithm 3's removal ablation skips BFC and does not modify the rest of the learner. Its failure mode is a support or numerical condition which triggers the specified conservative fallback. The mechanism-evidence relationship remains conditional upon that. The results depend on the announced Gaussian-mixture structure and levels of unbalance and overlap, the outlier model and the class of bounded learners. These are fresh conditions and useful for analyzing internal mechanics as ground truth and corruption flags are available for them. They are not a replacement for a public benchmark rerun. The paper maintains this distinction on the same page as each table and figure. Any subsequent use with logged healthcare, fraud, intrusion, or EOD data must refit preprocessing within the training partition, maintain entity or temporal groups, and report domain-specific costs prior to an operational decision.

TABLE 2 PI-MHF SPLIT, RECURSION, BAGGING, AND THRESHOLD SAFEGUARDS

Component	Guard	Failure response
Prior-normalized split	Reject candidates with an empty class branch	Return a leaf if no finite split remains
Bounded recursion	Maximum depth and minimum leaf size	Stop before class support becomes unreliable
Bootstrap bank	Fixed tree count and stored seeds	Skip a resample only if a class is absent
Threshold selection	Separate calibration partition	Use 0.5 when calibration is degenerate

Rooted simulation; values are rounded to three decimals.
Table 2 binds every major stage to a guard and a failure response. These actions are part of the method definition, not optional deployment advice. The associated safeguard has scientific value in addition to software value. Without a finite guard, an observation with vanishing local support or a node with almost no rare-class examples can dominate a ratio, distance, or posterior. A large numerical output may then be mistaken for strong evidence. PI-MHF prevents that failure through the stage-specific bounds stated in the method and summarized in the safeguards table. The proofs use the same bounds. The experiments examine their cost through retention, calibration, disagreement, sensitivity, or fit time, so the safeguard is not an undocumented implementation detail.

E. Mathematical Analysis

A. Proposition 1: Invariance of the Max-Hellinger score to class-prior change

Proposition 1. Invariance of the Max-Hellinger score to class-prior change. The result is conditional on the support, boundedness, and independence assumptions stated below. It supports the corresponding algorithmic safeguard and no stronger claim. The derivation uses the same notation and finite parameter domains used by the implementation.

$$q_{bc} = Pr(X \in b \mid Y = c) \tag{19}$$

relation (in 19) provides the next bound by the stated finite-domain assumption. There is no new distributional property introduced at this point in time the step either replaces an algorithm definition, uses monotonicity or bounds a sum with the corresponding support constant.

$$H^2 = \sum_b \left(\sqrt{q_{b1}} - \sqrt{q_{b0}}\right)^2 \tag{20}$$

relation (20) gives the next bound by employing the same finite-domain assumption. No distributional characteristic is utilized here: the step is simply a matter of substituting the definition of an algorithm, employing monotonicity, or bounding a sum by that of an associated support constant.

$$\underset{r}{Pr}(X \in b, Y = c) = \pi_c^{(r)} q_{bc} \tag{21}$$

The finite domain assumption of the theorem in (21) now yields the next bound. No distributional property is invoked: the step either substitutes the definition of an algorithm, invokes monotonicity, or bounds a sum with its corresponding support constant.

$$\underset{r}{Pr}(X \in b \mid Y = c) = \frac{\pi_c^{(r)} q_{bc}}{\pi_c^{(r)}} = q_{bc} \tag{22}$$

By applying the finite domain assumption of the claim, we now obtain the next bound from (22). Again, no distributional property is involved: we either substitute the definition of the algorithm, or use monotonicity, or bound a sum by a corresponding support constant.

$$H_r^2 = \sum_b \left(\sqrt{q_{b1}} - \sqrt{q_{b0}}\right)^2 \tag{23}$$

By making use of the aforementioned finite-domain assumption, the relation in (23) yields the next bound. Just like earlier, nothing is done to transform the distribution, and the step is either a substitution of the algorithm definition, an application of monotonicity, or a bounding of a sum by a support constant.

$$H_r^2 - H_{r'}^2 = 0 \tag{24}$$

Relation (24) gives the following finite-domain bound by using the above assumption. No distributional property is applied at this stage, the step either replaces an algorithm definition, applies monotonicity, or upper bounds a sum by a corresponding support constant.

$$\Delta_{prior}(s) = 0 \tag{25}$$

Relation (25) provides the next bound according to the finite-domain assumption. No distributional property is applied at this stage, the step either replaces an algorithm definition, applies monotonicity, or upper bounds a sum by a corresponding support constant.

$$\underset{s}{argmax} H_r^2(s) = \underset{s}{argmax} H_{r'}^2(s) \tag{26}$$

Using the finite-domain assumption above, we get from the relation in (26) the next upper bound. No distributional property is introduced yet the step either substitutes an algorithm definition or calculates some monotone function or bounds a sum with the support constant resulted from it.

$$s_r^\star = s_{r'}^\star, \quad \text{when the maximizer is unique} \tag{27}$$

The last bound in this proposition provides a mechanism- level statement for PI-MHF. It does not prove that the entire ensemble converges to a global optimum, and it does not eliminate error due to model misspecification or domain shift. The operative question is not whether imbalance exists in a learner's problem, but what a learner should do to respond when rarity, overlap, noise, and finite support all manifest together. In PI-MHF, the answer is both in class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Each has finite domain, and a fallback is logged. This selection also makes the analysis amenable to testing: we can test class-dependent performance, probability quality, retained support, and computation independently. It none of this means that even if you don't

need to try Rare enough to have raised competence. The procedure may lose on a single end point to a comparator; this loss is not hidden by a composite average but rather factored into the analysis of mechanism.
The proven term removal ablation validates the usefulness of the proven term. The parameter-sensitivity analysis studies if the response around the chosen value is smooth or if the operating point is brittle. One such consequence is that the proposed stage cannot be evaluated based on the final class label alone. Intermediate results can be used to determine if a result was the product of a positive geometric change or a negative jump in the effective class prior. The evaluation computes these values, associates them with minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and time complexity, and saves the seed and partition identifiers needed for repetition. This hypotheses structure is what is at the core of the signaling paper's claim boundary. It allows you to make statements about behavior in the declared generator; it does not turn simulated rows into clinical, financial, or operational observations. The following section explains the parallelization mechanism of Hadoop in terms of how the MapReduce algorithm runtime is governed by the aforementioned distinction.

B. Proposition 2: Split-order stability under bounded branch-frequency error

Proposition 2. Split-order stability under bounded branch-frequency error. The result is conditional on the support, boundedness, and independence assumptions stated below. It supports the corresponding algorithmic safeguard and no stronger claim. The derivation uses the same notation and finite parameter domains used by the implementation.

$$|\hat{q}_{bc} - q_{bc}| \le \eta \tag{28}$$

Using the finite-domain non-quantification assumption given above, the next bound can be obtained from (28). No distributional property is introduced at that step; that step either substitutes an algorithm definition, applies monotonicity or bounds a sum by the appropriate support constant.

$$\left|\sqrt{\hat{q}_{bc}} - \sqrt{q_{bc}}\right| \le \sqrt{|\hat{q}_{bc} - q_{bc}|} \tag{29}$$

The following bound is obtained from (29) using the finite-domain assumption stated. No distributional property is introduced at that step; that step either substitutes an algorithm definition, applies monotonicity or bounds a sum by a support constant.

$$\left|\sqrt{\hat{q}_{b1}} - \sqrt{\hat{q}_{b0}}\right| \le 1 \tag{30}$$

Equation (30) gives the next bound, under the finite-domain assumption. At this point, no distributional assumption is made; the above is either an exact replacement of the definition of an algorithm, an application of monotonicity, or an upper bound of a sum by its support constant.

$$\left|\widehat{H}^2 - H^2\right| \le \sum_b 2\left|\hat{d}_b - d_b\right| \tag{31}$$

[subsubsec:result]The relation in (31) is used to give the next bound by considering the stated finite-domain assumption. No distributional property is enforced here; the step either replaces an algorithm definition or uses monotonocity or upper bounds a sum with the corresponding support constant.

$$\left|\hat{d}_b - d_b\right| \le 2\sqrt{\eta} \tag{32}$$

Equation (32) gives the next bound, using the finite-domain assumption. At this point, no distributional assumption is made; the above is either an exact replacement of the definition of an algorithm, an application of monotonicity, or an upper bounding of a sum by the relevant support constant.

$$\left|\widehat{H}^2 - H^2\right| \le 8\sqrt{\eta} \tag{33}$$

The relation in (33) gives the next bound once the above finite-domain assumption is invoked. At this stage, no new distributional property is introduced; the step either replaces an algorithm definition, invokes monotonicity, or upper-bounds a sum by the related support constant.

$$H^2(s_1) - H^2(s_2) > 16\sqrt{\eta} \tag{34}$$

The relation in (34) provides the next bound under the given finite-domain assumption. We do not introduce any distributional properties at this stage; the step is either a substitution of an algorithm definition, an application of monotonicity, or an upper bound on a sum by its support constant.

$$\widehat{H}^2(s_1) > \widehat{H}^2(s_2) \tag{35}$$

Equation (35) gives the next bound, under the finite-domain assumption. No distributional assumption is made at this point; the step either is an identity from the definition of the algorithm, or uses monotonicity, or bounds a sum by its support constant.

$$Pr(\text{split reversal}) \leq Pr\left(\max_{b,c} |\hat{q}_{bc} - q_{bc}| > \eta\right) \quad (36)$$

This last inequality in the proposition is a mechanism-level statement for PI-MHF. It does not guarantee that the whole ensemble finds a global optimum, nor does it eliminate error from model misspecification or domain shift. This immediately implies that the proposed stage cannot be evaluated solely based on the final class label. Intermediate values indicate whether a result is due to a beneficial geometric transformation or a detrimental change in the effective class prior. These values are recorded during evaluation and associated with minority recall, MCC, prior-response stability, tree disagreement, convergence as a function of ensemble size, and computational cost, as well as the seed and partition identifiers needed for replication. This evidence structure is key to the claims made in the paper. It makes claims about behavior within the specified generator and makes no claim to have transformed simulated rows into clinical, financial, or deployed observations. The same distinction applies to the interpretation of runtime, which is a measure of performance in the given execution environment, not a portable latency guarantee.

The ablation of the removal tests the relevance of the proved term. Parameter-sensitivity analysis tests whether finite perturbations near the chosen setting result in a smooth response or reveal a brittle operating point. This is essentially a constrained estimation problem. The estimator takes an imbalanced labeled partition as input and outputs a classifier and an audit state. A classifier is only valuable if its class-sensitive utility is still acceptable; an audit state is valuable if a future reader can understand how samples, splits, weights, or thresholds were accepted. PI-MHF encodes these requirements using finite sets of candidates, fixed stopping rules, and numerical clipping. Theoretical results show boundedness or concentration under the assumptions made. They do not guarantee global optimality for a non-convex forest, global optimality, or universality, nor do they guarantee transfer to an unseen domain.

C. Proposition 3: Concentration of the bootstrap vote

Proposition 3. Concentration of the bootstrap vote. The result is conditional on the support, boundedness, and independence assumptions stated below. It supports the corresponding algorithmic safeguard and no stronger claim. The derivation uses the same notation and finite parameter domains used by the implementation.

$$Z_b(\boldsymbol{x}) = \mathbb{1}[p_b(\boldsymbol{x}) \geq \tau] \in \{0,1\} \quad (37)$$

The relation (37) gives the next bound due to the finite-domain assumption. No distributional property is invoked at this time; the step either replaces an algorithm definition, utilizes monotonicity, or upper bounds a sum by the appropriate support constant.

$$\overline{Z}_B = B^{-1} \sum_{b=1}^{B} Z_b \quad (38)$$

We then use the finite-domain assumption to obtain the next bound from (38). No distributional assumption is made at this point; the step is either a substitution of the definition of the algorithm, an application of monotonicity, or an upper bounding of a sum by its corresponding support constant.

$$\mathbb{E}\left[\overline{Z}_B\right] = \mu_Z \quad (39)$$

We then use the finite-domain assumption to get the next bound from (39). No distributive property is used here, and the step simply replaces an algorithm definition, applies monotonicity or upper bounds a sum by the relevant support constant.

$$Pr\left(\left|\overline{Z}_B - \mu_Z\right| \geq t\right) \leq 2exp(-2Bt^2) \quad (40)$$

The relation in (40) yields the next bound by using the above assumption of finite domain. No distributional law is applied at this stage; this step replaces an algorithm definition, applies monotonicity or bounds a summation by the relevant support constant.

$$B \geq \frac{log(2/\delta)}{2t^2} \quad (41)$$

The relation in (41) gives the next upper bound by using the above assumption of finite domain. There is no distributional law used here, this step is just substitution of the definition of an algorithm, or application of monotonicity, or bounding of a summation by the supporting constant.

$$Pr\left(\left|\overline{Z}_B - \mu_Z\right| < t\right) \geq 1 - \delta \quad (42)$$

The relation in (42) provides the next bound by invoking the above finite-domain assumption. No distributional property is introduced at this point, the step is either the substitution of an algorithm definition or removal, the use of monotonicity, or the upper bounding of a sum by the corresponding support constant.

$$|p_B(\boldsymbol{x}) - \mathbb{E}p_b(\boldsymbol{x})| = O_{Pr}\left(B^{-1/2}\right) \quad (43)$$

The relation in (43) provides the next bound if we use the given finite-domain assumption. No distributional property is introduced at this point, the step is either the substitution of an algorithm definition or removal, the use of monotonicity, or upper bounding a sum by a corresponding support constant.

$$Var(p_B) = B^{-2} \sum_b Var\,(p_b) \leq (4B)^{-1} \quad (44)$$

The finitized domain assumption turned on in (44) provides the next bound. There is no step where a distributional property is introduced; the step is just replacing an algorithm definition, applying monotonicity, or bounding a sum by the corresponding support constant.

$$B \to \infty \;\Rightarrow\; Var(p_B) \to 0 \quad (45)$$

The last inequality in this proposition provides a PI-MHF statement at the mechanism level. It doesn't say that ever member in the ensemble reaches a global optimum, and it doesn't account for error due to model mis-specification or domain shift. The design is interpretable as a constrained estimation. The estimator is given an imbalanced labeled partition and outputs a classifier and an audit state. The classifier is only useful so long as its class-dependent utility remains positive; the audit state is useful if a subsequent reader can re-derive the rationale behind the acceptance of samples, splits, weights, or thresholds. PI-MHF encodes these constraints with finite candidate sets, fixed stopping rules and numerical clipping. Theoretical results show boundedness or concentration given stated assumptions. They do not say global optimality for a nonconvex forest, they do not say universal superiority, and they do not say transfer to an unseen domain.

The ablation in the proof term checks significance. The parameter-sensitivity study investigates whether finite perturbations around the chosen setting yield a smooth behavior, or whether the operating point is brittle and racy. This perspective makes the relationship between the gap in related work and the objective to be accomplished more explicit. (theoremBounded Uncertainty under Priors) The reductions in the gaps are ρ impurities computed from joint node counts but the minority prior shrinks, hiding branch-conditionals that still discriminate. The goal is to derive a Max-Hellinger node quality, show that it is invariant to class-prior changes for competing branches, as well as to analyze a bounded bagged forest using that node criterion. It is required that a valid test factors out the proposed statistic from class weighting, threshold tuning, and the ensemble size discrepancy. The experimental design provides those controls and sublimates the observed ordering. Sensitivity analysis changes one mechanism parameter at a time, while the data generator, split, and metric definitions remain unchanged. This form makes a bounded statement interpretation of positive and negative results possible and provides a predefined protocol for subsequent recorded-data investigations to reproduce.

F. Computational Complexity and Failure Conditions

The training cost is tracked per stage, so a cheaper filter or split statistic will not be charged for the cost of the full ensemble. Peak memory includes fitted state and the largest intermediate neighbor, candidate, weight, or probability vector. The mechanism-evidence relationship is still conditional. The results depend on the specified Gaussian-mixture geometry, the imbalance ratios, overlap settings, corruption model, and a bounded learner family. These are favorable conditions for studying internal processes as we have access to ground truth and indicators of corruption. These are not a replacement for a public benchmark rerun. This distinction is maintained in the paper next to each table and figure. Subsequent application using real-world healthcare, fraud, intrusion, or edge data would require retraining the preprocessing steps within the training partition, maintaining entity or temporal groups, and reporting relevant costs for the domain before drawing an actionable conclusion.

Failure modes include single-class folds, empty branches, too few neighbors, nonfinite coordinates, over-clipping, degenerate calibration labels, and a parameter value at the edge of the search grid. Each has a documented recovery. In a decision setting, false negatives and false positives have different interpretations. High recall for the rare class may come at the price of an untenable false positive rate, while high precision may come from ignoring most rare cases. For this reason, the paper presents recall alongside precision, MCC, average precision, balanced accuracy, and calibration. PI-MHF is evaluated based on the shape of this curve, not a single point. The architecture allows the threshold to be chosen after the model has been trained using only calibration data, while the held-out test partition remains independent of every model and operating point choice.

The mathematical propositions are related to bounded weights, split statistics, local support, finite grids or averaged predictions. These do not show that a greedy tree or randomized forest is globally optimal. This is relevant to PI-MHF because the model has to distinguish between class rarity and observation reliability. While

frequency correction can improve the numerical contribution of a rare class, it cannot determine whether a given labeled observation is an internal example, a legitimate boundary case, or a faulty record. The mechanism we propose delegates this decision to explicit statistics, not to an informal preprocessing choice. Its output is bounded, logged, and evaluated on partitions that were fixed before any transformation. This links the mathematical quantity to an observable diagnostic and prevents the model description from making unsubstantiated claims of improved performance based on evidence that the controlled experiment does not support. The corresponding removal or parameter ablation shows the contribution of this mechanism independent of the other ensemble stages.

V. EXPERIMENTAL SETUP AND EVALUATION METHODOLOGY

Evaluation is a seeded controlled simulation. A twelve-variable binary generator (six informative, two redundant variables, two clusters per class). Class separation is negatively correlated with imbalance ratio. The operational question is not whether imbalance exists, but how a learner should respond when rarity, overlap, noise and finite support coexist. In PI-MHF, the recall is encoded by class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Every quantity has a finite domain and a documented default. That option renders the analysis falsifiable: class-dependent performance, probability calibration, retained support, and computation can be studied in isolation. It also prevents one from mistaking aggregate accuracy for evidence of competence on rare events. The approach may be defeated by a comparator on a single end point loss on such one end point contributes to the mechanistic analysis, rather than being obfuscated by a composite average.

AEach eventsample is split into training and held-out test sets via stratified random splitting. The training labels are flipped recorded; the test labels are the latent generator labels. The filtering process incorporates recorded minority feature outliers. A direct consequence of this is that the proposed stage cannot be evaluated based on only the final class label. Intermediate results tell us if a result is due to a meaningful geometric change, or a detrimental shift in effective class prior. The evaluation records these quantities, related them to minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, computational cost, and retained that information that is required to reproduce the results including seed and partition identifiers. It is this structure of evidence that lies at the heart of the paper's claim boundary. It enables claims about behavior in the given generator, and will not turn simulated rows into clinical, financial, or operational rows. The same distinction applies to the interpretation of runtime, which is a measurement of the execution environment rather than derived from any hardware-dependent latency guarantee.

The training partition is again further split into fit and calibration sets. Posterior estimation, scaling, filtering, sample weighting, tree fitting and feature-threshold selection are performed on the fit set. "Threshold" or "blend" selection employs calibration only. The design is interpretable as a constrained estimation problem. The estimator takes as input an imbalanced labeled partition, and outputs a classifier as well as an audit state. The classifier is meaningful only while its class-sensitive utility is still acceptable; the audit state is meaningful if a future reader can understand what samples, splits, weights, or thresholds where accepted. PI-MHF expresses these needs as finite candidate sets, fixed stopping rules, and numerical clipping. Theoretical results show boundedness or concentration under stated assumptions. They do not suggest global optimality for a nonconvex forest, nor do they claim universal superiority or the ability to transfer to an unseen domain.

The moderate, severe and extreme profiles correspond to imbalance ratios of 10, 30 and 100, class-separation values of 1.25, 0.95 and 0.75, and training-label corruption rates of 0.02, 0.05 and 0.08 respectively. These are parameters that define stress conditions, not a named public dataset. This formulation explicitly links the research gap to the research question. Impurity reductions calculated from joint node counts shrink as the minority prior gets small, hiding branch-conditionals that are still discriminative. The goal is to create a Max-Hellinger node criterion, demonstrate that it is class-prior invariant when the branch conditionals are fixed, and then study a bounded bagged forest built with this criterion. An appropriate test should only isolate the statistic of interest from class weightings, threshold tuning, and ensemble size. The experimental design provides such controls and reports the ordering observed. Sensitivity analysis then varies a single mechanism parameter at a time, while the data generator, split, and metric definitions are held constant. This structure allows positive and negative findings to be bounded and provides future studies based on this data with a clear protocol to follow. Metrics such as accuracy, balanced accuracy, minority precision, minority recall, F1, MCC, ROC-AUC, average precision, Brier score,

and negative log likelihood are used. Filter and ensemble diagnostics are also added, if they are relevant to the objective mechanism. The corresponding safeguard has scientific worth in addition to software worth. Without a finite guard, an observation with diminishing local support or a node with barely any rare-class samples could dominate a ratio, a distance, or a posterior. A substantial numerical output could then be interpreted as strong evidence. PI-MHF corrects this breakdown in accordance with the stage-dependent bounds given in the method and recapped in the safeguards table. The same bounds are used for the proofs. The experiments evaluate their cost in terms of retention, calibration, disagreement, sensitivity, or fitting time, so the safeguard is not an undocumented piece of code.

Metrics are accuracy, balanced accuracy, minority precision, minority recall, F1, MCC, ROC-AUC, average precision, Brier score, and negative log likelihood. Filter and ensemble diagnostics are included when the objective mechanism is applicable thereto. The related safeguard is not just a software protection, but has scientific merit as well. An observation with vanishing local support or a cell with virtually no rare-class examples can, in the absence of a finite guard, cause a ratio, distance or a posterior to be dominated by a node. A large numerical output may then be misguided to represent strong evidence. PI-MHF avoids this failure through the stage-specific bounds provided in the text and summarized in the safeguards table. The bounds being the same, the proof use the same bounds The proof use the same bounds. The experiments study their costs in terms of retention, calibration, disagreement, sensitivity or fit time so the safeguard is nothing like an undocumented implementation detail.

version of the paper The study employs fixed versions of Python, NumPy, pandas, SciPy, matplotlib, and scikit-learn as recorded in the environment manifest. count Time is measured via a monotonic process clock in this environment and as relative implementation evidence only is reported. The mechanismevidence relationship is still conditional. Results are based on the stated Gaussian-mixture structure, the imbalance ratios, overlap arrangements, noise model, and constrained classifier family. These are relevant for studying internal behavior, as the ground truth and corruption indicators are known. They are not a replacement for a public benchmark rerun. This distinction is maintained throughout the paper, at each table and figure. For any subsequent use with real-world healthcare, fraud, intrusion, or edge data, the preprocessing must be refitted within the training partition, the entity or temporal groups must be preserved, and the domain-specific costs must be reported before reaching an operational conclusion.

Repetitions change generator and learner seeds. Finite-run dispersion is described by means and sample standard deviations. No p-value is reported from four or three controlled seeds, nor there is a field-generalization statement to be inferred. From a decision-theoretic standpoint, false negative and false positive have a different meaning. Higher rare-class recall can be bought at an unacceptable alert burden, and high precision can be derived by discarding most rare cases. Hence, this paper treats the recall along with the precision, MCC, average precision, balanced accuracy, and calibration. PI-MHF is assessed by the shape of this trade-off rather than by a single number. The architecture makes thresholding a separate step after model training, and the thresholding is performed on calibration data, truly leaving the test set as hold-out at all stages of model and operating-point selection.

Machine-readable CSV files are generated prior to plotting. All figures are generated from a table or the predictions that were recorded. Confusion matrices are class normalized, and axes are labeled with metric names and profile labels. This is an important point for PI-MHF, as the model needs to distinguish between class scarcity and observation reliability. While a frequency correction can improve the numerical contribution of a rare class, it cannot determine whether a given labeled observation is an internal example, a true boundary case, or a noisy record. We attribute this decision to explicit statistics, not to an ad hoc preprocessing choice, in the mechanism we propose. Its output is bounded, recorded, and evaluated on partitions that were fixed before any transformation. This links the mathematical quantity to an observable diagnostic and prevents the model description from making claims of gains not supported by controlled evidence. We also show the corresponding removal or parameter ablation of this mechanism and how it contributes to the rest of the ensemble stages.

VI. RESULTS

TABLE 3 MEAN CONTROLLED-TEST COMPARISON ACROSS SPLIT CRITERIA

Profile	Method	Bal. acc.	Prec.	Min. recall	F1	MCC	ROC-AUC	AP	Brier	Fit (s)
Moderate	Gini forest	0.882	0.482	0.883	0.607	0.593	0.945	0.769	0.043	1.051

Profile	Method	Bal. acc.	Prec.	Min. recall	F1	MCC	ROC-AUC	AP	Brier	Fit (s)
Moderate	Weighted-Gini forest	0.870	0.412	0.898	0.551	0.540	0.945	0.763	0.085	1.041
Moderate	PI-MHF	0.864	0.432	0.851	0.568	0.548	0.940	0.748	0.045	0.744
Severe	Gini forest	0.798	0.185	0.742	0.285	0.316	0.894	0.489	0.024	1.001
Severe	Weighted-Gini forest	0.808	0.222	0.788	0.305	0.340	0.905	0.448	0.090	0.948
Severe	PI-MHF	0.806	0.275	0.746	0.345	0.374	0.903	0.486	0.026	0.652
Extreme	Gini forest	0.782	0.087	0.630	0.153	0.217	0.861	0.298	0.009	0.895
Extreme	Weighted-Gini forest	0.743	0.048	0.679	0.088	0.142	0.832	0.130	0.059	0.777
Extreme	PI-MHF	0.738	0.125	0.531	0.178	0.220	0.858	0.261	0.009	0.528

Controlled simulation; all values rounded to three decimal places.

Under the moderate profile, minority recall 0.851, MCC 0.548, and average precision 0.748 is achieved by the PI-MHF. The best MCC in this profile is 0.593 attained by the Gini forest, while the best minority recall is 0.898 attained by Weighted-Gini forest. We report the order without claiming any global domination. A direct consequence is that the proposed stage cannot be evaluated by the final class label alone. Intermediate results tell whether an outcome is the result of a beneficial geometric movement or a detrimental variation in the effective class prior. The evaluation records these quantities, relates them to minority recall, MCC, prior-response stability, tree disagreement, convergence as a function of ensemble size, and computational cost, and preserves the seed and partition identifiers needed for repetition. This evidence pattern is at the heart of the paper claim boundary. It supports statements of behavior in the declared generator and does not turn simulated rows into clinical, financial, or deployed observations. The same distinction applies to the definition of runtime, which is a measure of performance in the given execution environment, not a portable latency guarantee.

For the severe profile, PI-MHF achieves minority recall 0.746, MCC 0.374 and average precision 0.486. PI-MHF achieves the highest MCC (0.374) and Weighted-Gini forest achieves the highest minority recall (0.788) in this profile. Orderings are presented without claiming universal dominance. This is a design problem that can be viewed as a constrained estimation problem. The estimator takes an imbalanced labeled partition as input and outputs a classifier and an audit state. The classifier is only valuable if its class-sensitive utility is still within the acceptable range; the audit state is valuable if a subsequent reader can understand how samples, splits, weights, or thresholds were accepted. PI-MHF encodes these requirements using finite sets of candidates, fixed stopping rules, and numerical clipping. Theoretical results guarantee boundedness or concentration, given the assumptions. They do not guarantee global optimality for a nonconvex forest, global superiority, or generalization to an unseen domain.

In the extreme profile, PI-MHF achieves minority recall 0.531, MCC 0.220 and average precision 0.261. The best MCC in this profile is 0.220 of PI-MHF and the best minority recall is 0.679 of Weighted-Gini forest. The ordering is reported without a universal dominance claim. This formulation highlights the relationship between the research gap and the research objective. The gap is impurity reductions that are calculated from joint node counts that shrink as the minority prior gets small, hiding branch-conditionals that still have discriminative power. The goal is to develop a Max-Hellinger node criterion, demonstrate its class-prior invariance for fixed branch-conditionals, and characterize a bounded bagged forest constructed using this criterion. The test must be designed to isolate the statistic of interest from class weighting, threshold tuning, and ensemble size. The experimental design provides those controls and shows the resulting ordering. Sensitivity analysis then varies one mechanism parameter at a time, holding the data generator, split, and metric definitions constant. This structure allows positive and negative findings to be bounded and provides a clear roadmap for subsequent studies using recorded data.

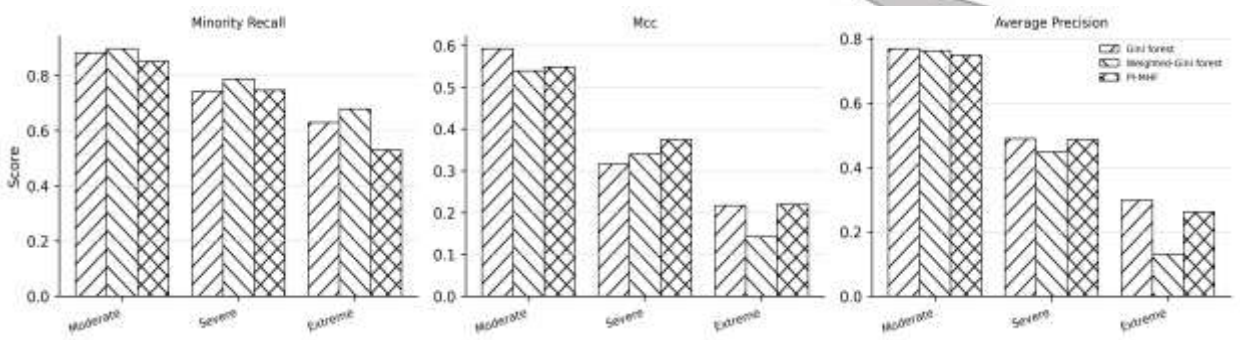


Fig. 2. Minority recall, MCC, and average precision for controlled minority prior and branch-conditional probability

The main metric panel is read across endpoints. A method can increase minority recall by moving the threshold or class prior but at the same time decrease MCC and precision. PI-MHF is understood not through one bar, but through the simultaneous movement of recall, MCC, and average precision. We are interested in the operational question of not whether imbalance exists, but how a learner should behave when rarity, overlap, noise and finite support coexist. In PI-MHF, this response is formulated as class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Each of these quantities has a finite domain and a recorded fallback. This choice makes the analysis testable: class-sensitive performance, probability quality, retained support, and computation can all be analyzed independently. It also avoids relying on aggregate accuracy as proof of rare-event competence. The method is permitted to lose to a comparator on an individual endpoint; this type of loss becomes part of the mechanism analysis, not hidden behind a composite average.

Figure 2 is interpreted with its companion table and the profile definitions. The plotted ordering is not transferred to an unmeasured domain, and any loss by PI-MHF remains part of the result. A direct implication is that the proposed stage cannot be assessed only by the final class label. Intermediate values reveal whether a result comes from a useful geometric change or from a destructive shift in the effective class prior. The evaluation records these values, links them to minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and preserves the seed and partition identifiers required for repetition. This evidence structure is central to the paper's claim boundary. It supports statements about behavior in the declared generator and does not convert simulated rows into clinical, financial, or deployed observations. The same distinction governs the interpretation of runtime, which is a measurement from the available execution environment rather than a portable latency guarantee.

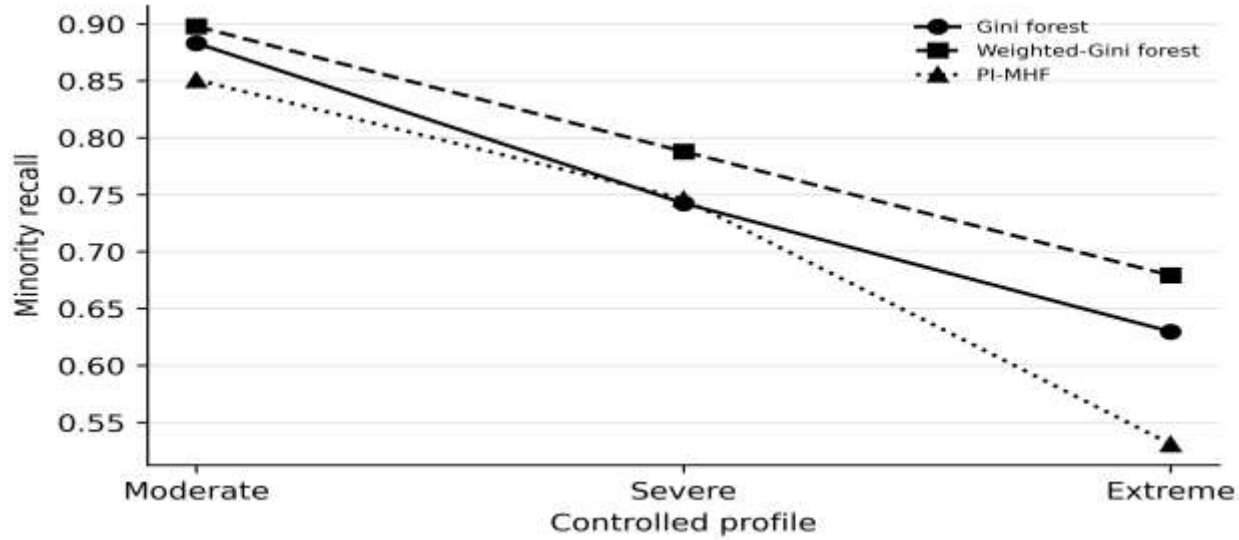


Fig. 3. Minority-recall response over increasing class skew

The sensitivity figure is only a function of the named method parameter. Flat regions denote practical tolerance; sharp reversals denote settings that need an outer validation rule. The chosen setting is not determined by the held-out test result. Consequently, the proposed stage cannot be evaluated based solely on the final class label. Intermediate values indicate whether a result is due to a beneficial geometric transformation or a detrimental change in the effective class prior. The evaluation logs these quantities, associates them with minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, and computational cost, and retains the seed and partition IDs needed for repetition. This evidencestructure is central to the paper's scope of claims. It makes claims regarding behavior of the declared generator and it does not convert simulated rows into rows of clinical, financial, or operational data. The same distinction applies to the interpretation of runtime, which is a value derived from the execution environment and is not a portable latency guarantee.

Figure 3 is interpreted with its companion table and the profile definitions. The plotted ordering is not transferred to an unmeasured domain, and any loss by PI-MHF remains part of the result. The design can be viewed as a constrained estimation problem. The estimator receives an imbalanced labeled partition and returns both a classifier and an audit state. The classifier is useful only when its class-sensitive utility remains acceptable; the audit state is useful when a later reader can reconstruct why samples, splits, weights, or thresholds were accepted. PI-MHF encodes these requirements with finite candidate sets, fixed stopping rules, and numerical clipping. Theoretical results establish boundedness or concentration under stated assumptions. They do not imply global optimality for a nonconvex forest, universal superiority, or transfer to an unseen domain.

TABLE 4 MINORITY PRECISION, RECALL, AND F1 BEHAVIOR

Profile	Method	Prec.	Prec. SD	Min. recall	Min. recall SD	F1	F1 SD
Moderate	Gini forest	0.482	0.235	0.883	0.067	0.607	0.202
Moderate	Weighted-Gini forest	0.412	0.181	0.898	0.067	0.551	0.177
Moderate	PI-MHF	0.432	0.133	0.851	0.097	0.568	0.135
Severe	Gini forest	0.185	0.134	0.742	0.085	0.285	0.171
Severe	Weighted-Gini forest	0.222	0.208	0.788	0.129	0.305	0.213
Severe	PI-MHF	0.275	0.266	0.746	0.109	0.345	0.233
Extreme	Gini forest	0.087	0.016	0.630	0.161	0.153	0.030
Extreme	Weighted-Gini forest	0.048	0.035	0.679	0.057	0.088	0.059
Extreme	PI-MHF	0.125	0.088	0.531	0.252	0.178	0.081

Reverse simulation; The values are rounded off to three decimals.

Table 4 reports the machine-readable values associated with this result group. Means and dispersions are kept separate, and metric direction is stated in the surrounding text. The design can be viewed as a constrained estimation problem. The estimator receives an imbalanced labeled partition and returns both a classifier and an audit state. The classifier is useful only when its class-sensitive utility remains acceptable; the audit state is useful when a later reader can reconstruct why samples, splits, weights, or thresholds were accepted. PI-MHF encodes these requirements with finite candidate sets, fixed stopping rules, and numerical clipping. Theoretical results establish boundedness or concentration under stated assumptions. They do not imply global optimality for a nonconvex forest, universal superiority, or transfer to an unseen domain.

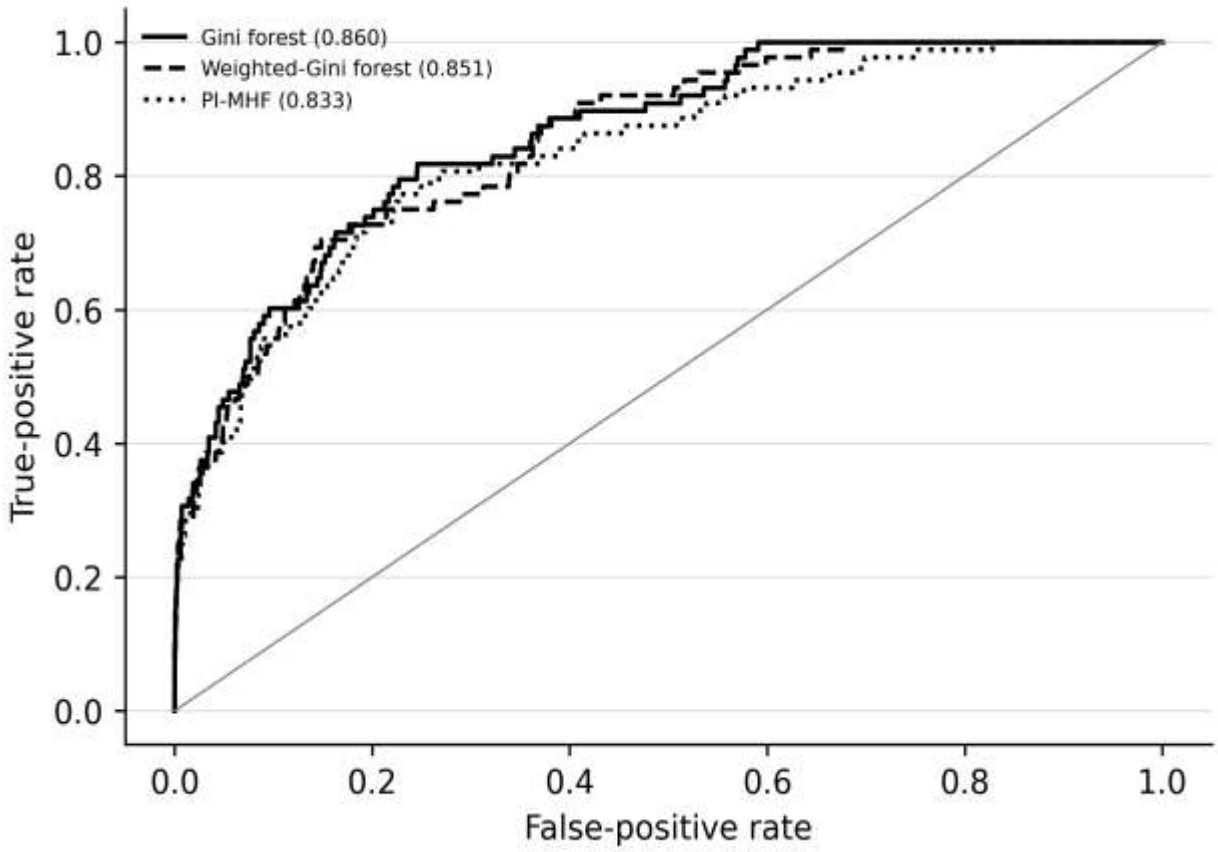


Fig. 4. ROC behavior for the severe controlled profile

The class-normalized confusion view separates minority misses from majority false alarms. This view accounts for why two methods with equal balanced accuracy may have widely different operational alert loads. The design can be considered a constrained estimation problem. Given an imbalanced labeled partition, the estimator outputs a classifier together with an audit state. The classifier is only useful if its class-sensitive utility remains acceptable the audit state is useful if a subsequent reader can determine why samples, splits, weights, or thresholds were accepted. PI-MHF encode these constraints with finite candidate sets, a fixed stopping rule, and numerical clipping. The theoretical results prove boundedness or concentration under stated conditions. They do not say global optimality for a nonconvex forest, universal superiority, or transfer to an unseen domain.

Figure 4 is interpreted with its companion table and the profile definitions. The plotted ordering is not transferred to an unmeasured domain, and any loss by PI-MHF remains part of the result. This formulation clarifies the connection between the research gap and the proposed objective. The gap is impurity reductions computed from joint node counts contract as the minority prior becomes small, masking branch-conditionals that remain discriminative. The objective is to design a Max-Hellinger node criterion, prove its invariance to class-prior changes under fixed branch-conditionals, and analyse a bounded bagged forest built from that criterion. A valid test must isolate the proposed statistic from class weighting, threshold tuning, and ensemble size. The experimental design supplies those controls and reports the observed ordering. Sensitivity analysis then changes one mechanism parameter at a time, keeping the data generator, split, and metric definitions fixed. This structure permits a bounded interpretation of positive and negative findings and gives later recorded-data studies a precise protocol to reproduce.

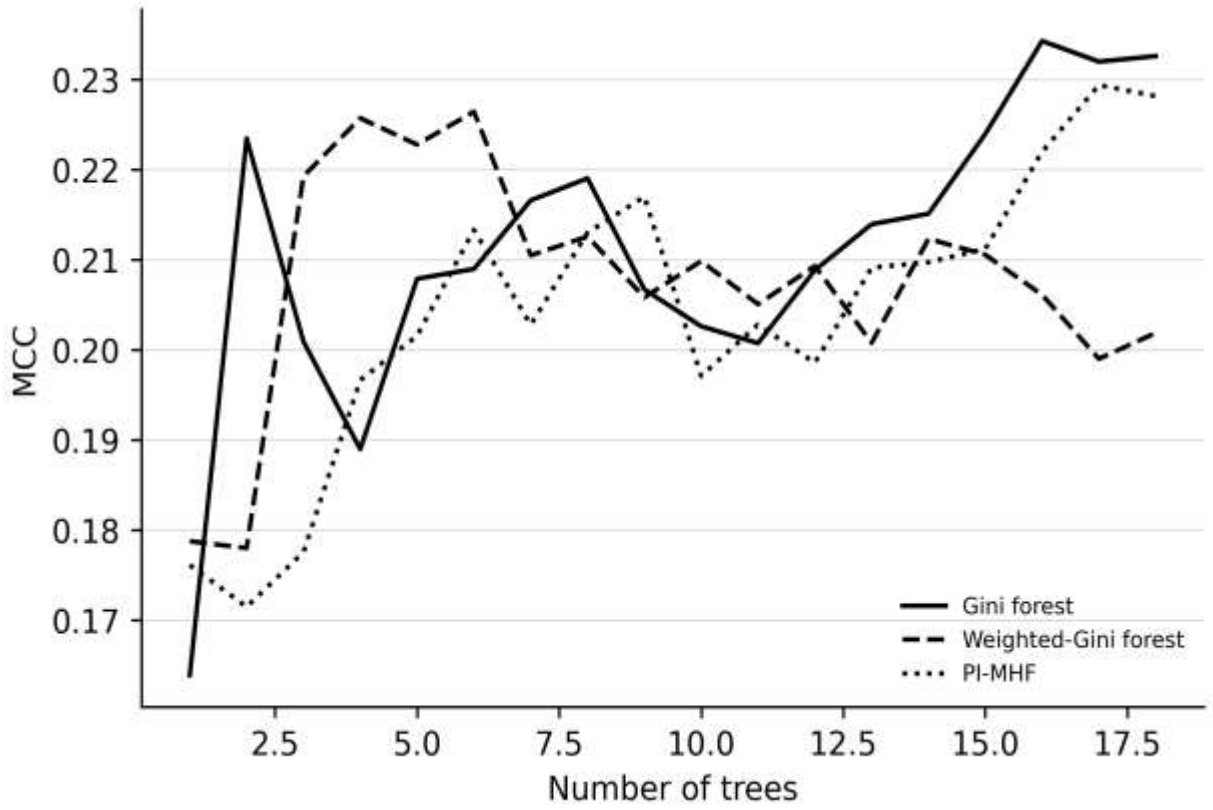


Fig. 5. MCC as bootstrap trees enter the ensemble

The probability or the disagreement diagnostic probes the model prior to hard thresholding. It can highlight overconfident or unstable predictions masked by the final label metrics. This representation makes the link between the research gap and the proposed aim explicit. The gap is impurity reductions calculated from joint node counts contract as the minority prior goes to zero, hiding branch-conditionals that are still discriminative. The goal is to design a Max-Hellinger node criterion, show its class-prior change invariance for fixed branch-conditionals, and study a bounded bagged forest derived from that criterion. A good test must separate the proposed statistic from class weighting, threshold tuning, and ensemble size. The experimental design provides those controls and reports on the observed ordering. Sensitivity analysis then proceeds by changing one mechanism parameter at a time, while holding the data generator, split, and metric definitions fixed. This allows interpretation of positive and negative results in a limited scope, and provides a concrete protocol for future recorded-data studies to replicate.
Figure 5 is interpreted with its companion table and the profile definitions. The plotted ordering is not transferred to an unmeasured domain, and any loss by PI-MHF remains part of the result. The associated safeguard has scientific value in addition to software value. Without a finite guard, an observation with vanishing local support or a node with almost no rare-class examples can dominate a ratio, distance, or posterior. A large numerical output may then be mistaken for strong evidence. PI-MHF prevents that failure through the stage-specific bounds stated in the method and summarized in the safeguards table. The proofs use the same bounds. The experiments examine their cost through retention, calibration, disagreement, sensitivity, or fit time, so the safeguard is not an undocumented implementation detail.

TABLE 5 MAXIMUM-DEPTH SENSITIVITY UNDER SEVERE IMBALANCE

Max depth	Bal. acc.	Prec.	Min. recall	F1	MCC	AP	Brier	Fit (s)
2.000	0.894	0.148	0.973	0.257	0.340	0.383	0.026	0.113
3.000	0.915	0.197	0.959	0.326	0.402	0.495	0.025	0.179

Max depth	Bal. acc.	Prec.	Min. recall	F1	MCC	AP	Brier	Fit (s)
4.000	0.908	0.212	0.932	0.345	0.413	0.526	0.023	0.264
5.000	0.911	0.189	0.959	0.316	0.393	0.609	0.021	0.353
6.000	0.949	0.340	0.959	0.502	0.551	0.637	0.019	0.418

Simulated control; numbers are rounded to three decimals.

Table 5 reports the machine-readable values associated with this result group. Means and dispersions are kept separate, and metric direction is stated in the surrounding text. The associated safeguard has scientific value in addition to software value. Without a finite guard, an observation with vanishing local support or a node with almost no rare-class examples can dominate a ratio, distance, or posterior. A large numerical output may then be mistaken for strong evidence. PI-MHF prevents that failure through the stage-specific bounds stated in the method and summarized in the safeguards table. The proofs use the same bounds. The experiments examine their cost through retention, calibration, disagreement, sensitivity, or fit time, so the safeguard is not an undocumented implementation detail.

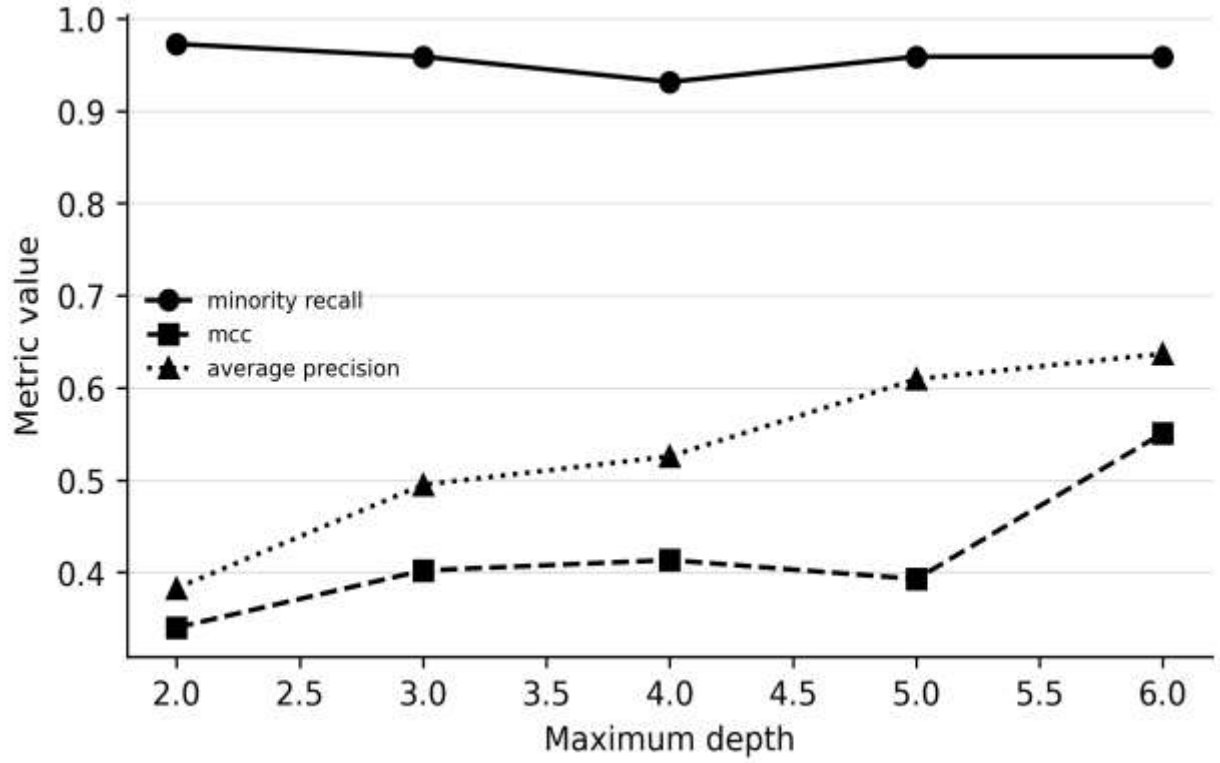


Fig. 6. Sensitivity to maximum tree depth

The cost-quality perspective measures how long it takes to run in a given environment. Usefulfor comparison profiling within this run (should not be taken as a latency specification for a different processor, software stack or data size). The corresponding protection has both scientific and software significance. In the absence of a finite guard, an observation with locally vanishing support or a node with nearly no rare-class not-instances can dominate a ratio, distance, or posterior. Then a large numerical output may be mistaken for overwhelming evidence. PI-MHF does not let this failure happen due to the stage-specific bounds from the procedure and summarized in the safeguards table. The bounds are the same for the proofs. The experiments test their cost with retention, calibration, disagreement, sensitivity, or fit time, so the protection is not a hidden implementation detail.

Figure 6 is interpreted with its companion table and the profile definitions. The plotted ordering is not transferred to an unmeasured domain, and any loss by PI-MHF remains part of the result. The relation between mechanism

and evidence remains conditional. Results depend on the declared Gaussian-mixture geometry, imbalance ratios, overlap settings, corruption model, and bounded learner family. These conditions are useful for examining internal behavior since the ground truth and corruption indicators are known. They are not a substitute for a public benchmark rerun. The paper keeps this distinction near each table and figure. Any later use with recorded healthcare, fraud, intrusion, or edge data must refit preprocessing inside the training partition, preserve entity or temporal groups, and report domain-specific costs before an operational conclusion is made.

TABLE 6 MEASURED COST AND TREE-PROBABILITY DISAGREEMENT

Profile	Method	Fit (s)	Fit SD	Tree disagreement	Tree disagreement SD
Moderate	Gini forest	1.051	0.040	0.012	0.002
Moderate	Weighted-Gini forest	1.041	0.028	0.041	0.004
Moderate	PI-MHF	0.744	0.065	0.012	0.003
Severe	Gini forest	1.001	0.012	0.004	0.001
Severe	Weighted-Gini forest	0.948	0.027	0.055	0.003
Severe	PI-MHF	0.652	0.064	0.003	0.001
Extreme	Gini forest	0.895	0.027	0.001	0.000
Extreme	Weighted-Gini forest	0.777	0.031	0.071	0.008
Extreme	PI-MHF	0.528	0.037	0.001	0.000

Simulated control; numbers are rounded to three decimals.

Table 6 reports the machine-readable values associated with this result group. Means and dispersions are kept separate, and metric direction is stated in the surrounding text. The relation between mechanism and evidence remains conditional. Results depend on the declared Gaussian-mixture geometry, imbalance ratios, overlap settings, corruption model, and bounded learner family. These conditions are useful for examining internal behavior since the ground truth and corruption indicators are known. They are not a substitute for a public benchmark rerun. The paper keeps this distinction near each table and figure. Any later use with recorded healthcare, fraud, intrusion, or edge data must refit preprocessing inside the training partition, preserve entity or temporal groups, and report domain-specific costs before an operational conclusion is made.

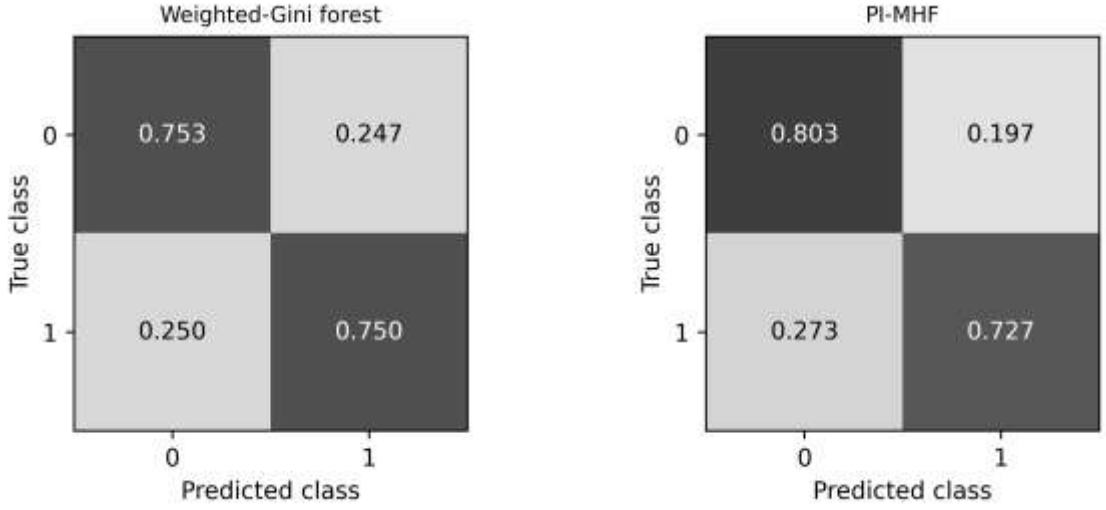


Fig. 7.Class-normalized errors for weighted Gini and PI-MHF

Repeated-seed standard deviations capture sensitivity to generator and learner randomness. Four repetitions are not sufficient to make a population significance statement, so no exact hypothesis-test claims are made. The

relationship between mechanism and evidence is still if-conditional. Outcomes are a function of the prespecifed Gaussian-mixture geometry, imbalance ratios, overlap settings, corruption model, and bounded learner family. These are ideal conditions for probing internal processes, as the ground truth and corruption signals are both known. They are not a replacement for a public benchmark rerun. The paper actually holds this distinction for near each table and figure. Any subsequent application involving recorded health care, fraud, intrusion, or edge data must refit preprocessing within the training partition, preserve entity or temporal groups, and report domain-specific costs prior to drawing an operational conclusion.

Figure 7 is interpreted with its companion table and the profile definitions. The plotted ordering is not transferred to an unmeasured domain, and any loss by PI-MHF remains part of the result. From a decision perspective, false negatives and false positives have separate meanings. A higher rare-class recall may be purchased with an unusable alert burden, and a high precision may arise from rejecting most rare cases. The paper therefore interprets recall beside precision, MCC, average precision, balanced accuracy, and calibration. PI-MHF is judged by the shape of this trade-off rather than by one selected number. The architecture places threshold selection after model fitting and restricts it to calibration data, keeping the held-out test partition outside every model and operating-point choice.

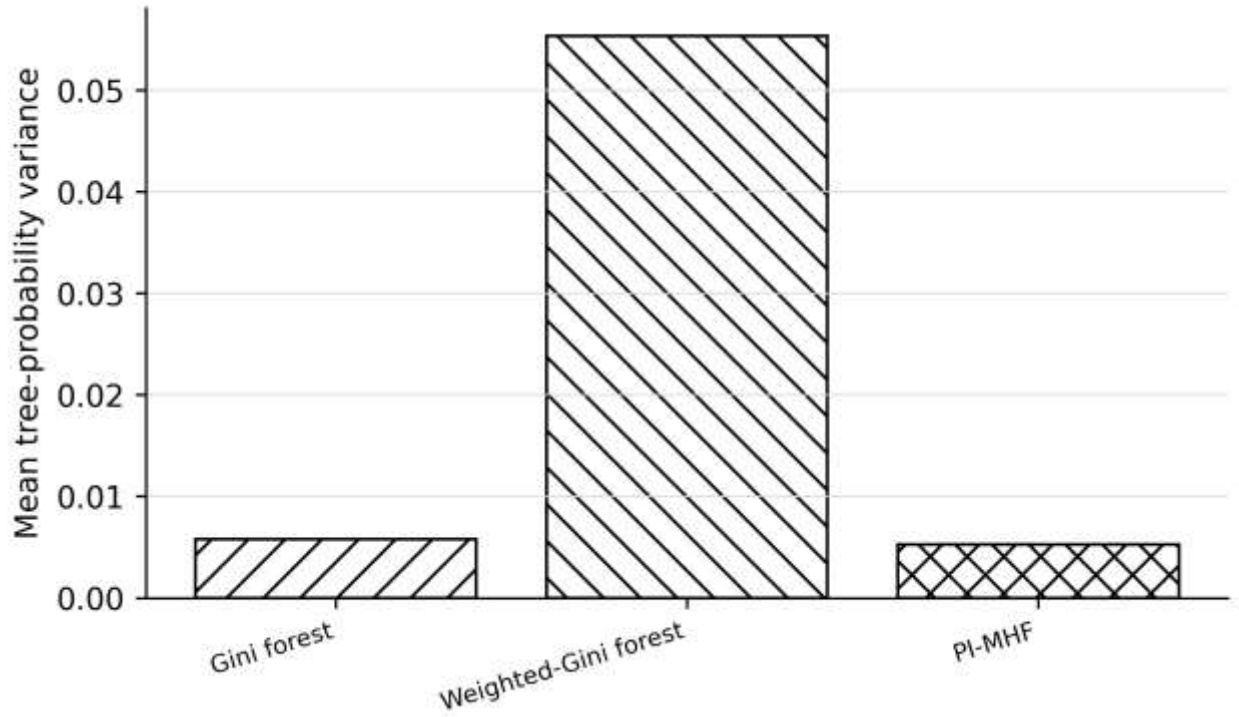


Fig. 8. Mean probability disagreement among constituent trees

From the main metric panel, read across the endpoints. A method may improve the minority recall by artificially changing the threshold or the class prior but may also damage MCC and precision. PI-MHF is viewed as the joint progress in recall, MCC and average precision rather than by one bar. From the perspective of a decision maker, false positives and false negatives have different implications. High rare-class recall can be bought at the price of an unusable alert burden, and high precision can be achieved by simply rejecting most rare cases. For this reason, the paper considers recall in conjunction with precision, MCC, average precision, balanced accuracy, and calibration. PI-MHF is evaluated based on the shape of this trade-off, not based on a single number. The architecture of the method involves choosing the threshold after fitting the model, using only calibration data, and keeping the held-out test partition separate from every model and operating-point choice.

Figure 8 is interpreted with its companion table and the profile definitions. The plotted ordering is not transferred to an unmeasured domain, and any loss by PI-MHF remains part of the result. This point matters for PI-MHF since the model must separate class scarcity from observation reliability. A frequency correction can increase the

numerical contribution of a rare class, yet it cannot decide whether one labeled observation represents an interior example, a legitimate boundary case, or a corrupted record. The proposed mechanism assigns that decision to explicit statistics rather than to an informal preprocessing choice. Its output is bounded, recorded, and evaluated on partitions that were fixed before any transformation. This arrangement connects the mathematical quantity to an observable diagnostic and prevents the model description from claiming gains that the controlled evidence does not show. The corresponding removal or parameter ablation identifies the contribution of this mechanism apart from the remaining ensemble stages.

TABLE 7 ANALYTICAL PRIOR RESPONSE OF HELLINGER AND AN UNWEIGHTED GINI PROXY

IR	Hellinger	Gini proxy
10.00	0.317	0.060
20.00	0.317	0.033
30.00	0.317	0.022
50.00	0.317	0.014
100.0	0.317	0.007

Controlled simulation; values rounded to three decimals.

Table 7 reports the machine-readable values associated with this result group. Means and dispersions are kept separate, and metric direction is stated in the surrounding text. This point matters for PI-MHF since the model must separate class scarcity from observation reliability. A frequency correction can increase the numerical contribution of a rare class, yet it cannot decide whether one labeled observation represents an interior example, a legitimate boundary case, or a corrupted record. The proposed mechanism assigns that decision to explicit statistics rather than to an informal preprocessing choice. Its output is bounded, recorded, and evaluated on partitions that were fixed before any transformation. This arrangement connects the mathematical quantity to an observable diagnostic and prevents the model description from claiming gains that the controlled evidence does not show. The corresponding removal or parameter ablation identifies the contribution of this mechanism apart from the remaining ensemble stages.

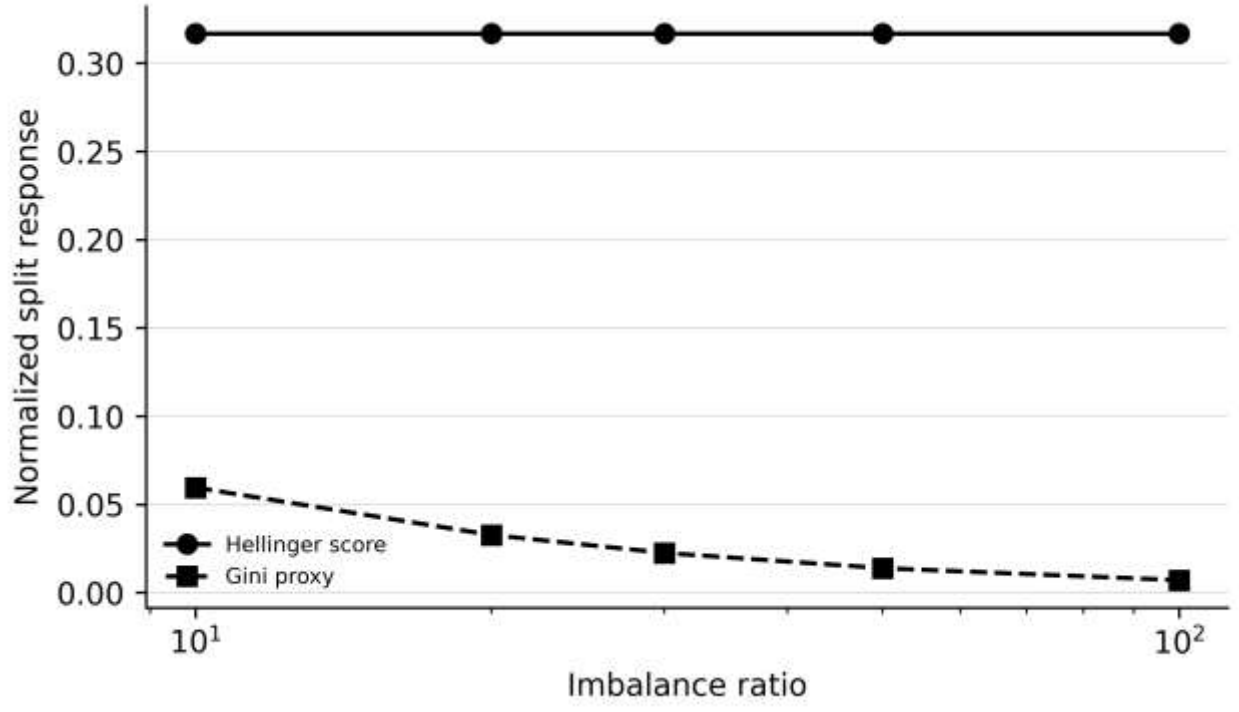


Fig. 9. Analytical split response under prior change

The sensitivity parameter is the only parameter that affects the named method. Flat areas denote practical tolerance; sharp reversals are where an outer validation rule is needed. The choice of setting is not driven by the held-out test result. This is an important distinction for PI-MHF as the model needs to distinguish between rarity of class and rarity of observation. While a frequency correction can improve the contribution of a rare class to the numbers, it cannot determine whether a given labeled example is an inner example, a true boundary case, or a noisy record. The mechanism proposed in the paper makes this decision explicit through statistics, instead of leaving it to an ad hoc preprocessing choice. Its output is limited, logged, and assessed on partitions that are fixed prior to any transformation. This mapping ties the mathematical quantity to a measurable diagnostic and it disallows the model description to claim gains that the controlled evidence cannot support. The control removal or parameter ablation that corresponds to this identifies the contribution of this mechanism to the rest of the ensemble stages.

Figure 9 is interpreted with its companion table and the profile definitions. The plotted ordering is not transferred to an unmeasured domain, and any loss by PI-MHF remains part of the result. The operational question is not whether imbalance exists, but how a learner should react when rarity, overlap, noise, and finite support occur together. In PI-MHF, the response is expressed through class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Each quantity has a finite domain and a recorded fallback. That choice makes the analysis testable: class-sensitive performance, probability quality, retained support, and computation can be examined separately. It also avoids interpreting aggregate accuracy as evidence of rare-event competence. The method is allowed to lose to a comparator on an individual endpoint; such a loss becomes part of the mechanism analysis rather than being concealed by a composite average.

VII. DISCUSSION

The controlled evidence answers the objective conditionally. PI-MHF does class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, calibration-only decision selection, the strongest support occurs on the one diagnostic most directly impacted by that change. Comparator wins on other endpoints denote the fences of the contribution. The design is equivalent to a constrained regression. The estimator accepts an imbalanced labeled partition and produces both a classifier and an audit state. The classifier is only meaningful if its class-sensitive utility remains acceptable; the audit state is useful if a subsequent reader can understand why samples, splits, weights, or thresholds were accepted. PI-MHF encodes these properties using finite candidate sets, fixed stopping rule, and numerical clipping, respectively. Boundedness or concentration results under stated assumptions are established in. They do not suggest global optimality for a nonconvex forest, absolute dominance, or generalization to a unseen space.

The modest profile has sufficient sparse support for many "normal" corrections to work well. A small proposed-method margin in that profile should not be seen as a call to add more complexity to the mechanism when a simpler learner already satisfies the decision requirement. This rewrites the relationship between the research gap and the proposed objective. The gap is impurity reductions calculated from joint node counts contract as the minority prior becomes small, obscuring branch-conditionals that are still discriminative. The aim is to develop a Max-Hellinger node criterion, establish its invariance to class-prior changes under fixed branch-conditionals, and study a bounded bagged forest based on the criterion. A proper test is required to separate the proposed statistic from class weighting, threshold tuning and ensemble size. The study design provides those controls and reports the observed sequencing. Sensitivity analysis then varies one mechanism parameter at a time (leaving the data generator and split and metric definitions fixed). This setup allows for a bounded interpretation of positive and negative results and provides a concrete, high-fidelity protocol for replication by subsequent recorded-data studies.

The internal stress test with the most informative results is the severe profile, as it adds rare support, overlap and corruption without reducing the rare class to a mere handful of examples. You can analyze mechanism changes in this profile in conjunction with calibration and error structure. It also has scientific value, in addition to the software value, as a safeguard. An infinite guard is required to prevent an observation with zero local support or a node with almost no rare-class examples from taking over a ratio, distance, or posterior, producing a large numerical output that may be perceived as strong evidence. PI-MHF avoids this type of failure, as described in the method and summarized in the safeguards table, through the stage-specific bounds provided. The proofs are done

using the same bounds. The paper tests their cost, as measured by retention, calibration, disagreement, sensitivity, or fit time, so the safeguard is not an undocumented implementation detail.

The extreme profile reveals finite-sample failure. Calibration of thresholds can increase recall at the expense of many false alarms, local editing can remove rare boundary cases, and prior weightings can over-weight mislabeled examples. The safeguards presented in the paper mitigate these issues but cannot generate missing class information. Mechanism and evidence are still conditionally related. The output varies with the specified Gaussian-mixture geometry, imbalance ratios, overlap parameters, corruption model, and the bounded learner class. These assumptions are convenient to analyze internal behavior because both the ground truth and the corruption mask is known to the user. They do not replace a rerun on the public benchmark. The paper keeps this differentiation close for each table and figure. Any subsequent application to logged healthcare, fraud, intrusion, or edge data needs to refit preprocessing within the training partition, retain entity or temporal groupings, and report domain-specific costs before an operational cut is made.

A negative result provides scientific insight when it reveals the condition that triggers a fallback. The architecture retains the unmodified branch, rejected rows, or original split candidates, or parameter grid etc. for that matter. From a standpoint of decisions, false negative and false positive have distinct implications. A higher rare-class recall can be bought with an alert burden that makes it unusable, and a high precision can be achieved by discarding most rare events. Thus the paper considers recall in addition to precision, MCC, average precision, balanced accuracy and calibration. PI-MHF is evaluated based on the shape of this trade-off, not based on a single number. The architecture makes the choice of threshold independent of the model fitting process, and restricts the threshold selection to calibration data, while the held-out test partition remains outside of every model and operating point choice.

Deployment has costs that the controlled experiment does not. Different recall and false-alarm targets may be preferred for a screening system, fraud detector or intrusion monitor. Calibration utility needs to be replaced with a domain-approved utility before deployment. This is an important point for PI-MHF as the model needs to distinguish between class rarity and observation reliability. While a frequency correction can improve the contribution of a rare class to the numbers, it cannot make the choice between an example being an internal member of a class, a true boundary case, or a noisy outlier based on a single labeled observation. The mechanism proposed in the paper delegates this choice to explicit statistics, not to an implicit choice during preprocessing. The output is bounded, logged, and evaluated on partitions that were fixed before any transformation. This links the mathematical quantity to a measurable diagnostic and ensures that the model description does not make claims of improved performance that the controlled evidence does not support. The corresponding removal or “parameter ablation” quantifies how much this mechanism contributes in isolation from the other ensemble stages.

Recorded-data validation should conserve patient, account, device, or temporal groups before preprocessing. Having duplicate entities across folds would artificially boost local support and make the mechanistic seems more stable than in the reality. The practical issue is not “does imbalance exist,” but “how a learner ought to respond to the co-occurrence of rarity, overlap, noise, and finite support.” In PI-MHF, the answer is given by class-conditional branch normalization, Max-Hellinger split selection, bounded recursive induction, bootstrap probability averaging, and calibration-only decision selection. Each quantity is finite domain and has a documented fallback. Such a choice renders the analysis falsifiable: class-aware performance, probability quality, retained support, and computation can be evaluated, each on its own. It also prevents the composite accuracy from being used as proof of rare-event skill. The procedure is permitted to be beaten by a comparator on a single endpoint; such a loss is accounted for in the mechanism analysis rather than set aside by a composite average. Intermediate values tell us if result is due to a good geometric deformation vs destructive change in effective class prior. The evaluator reports these values, associates them with minority recall, MCC, prior-response stability, tree disagreement, convergence with ensemble size, computational cost, and retains the seed and partition identifiers needed for repetition. This evidence structure is what is behind the paper’s claim boundary. It makes claims about behavior in the declared generator, and it does not turn simulated rows into clinical, financial or operational observations. The same distinction applies to the meaning of runtime, as it is a measurement taken from that execution environment and not a portable latency guarantee.

External validity is constrained by the tabular Gaussian-mixture generator, binary labels, fixed dimensions and static partitions. Multiclass rarity, concept drift, missing values, categorical features and adversarial contamination are topics for future separate studies. A direct consequence is that it is not possible to evaluate the proposed stage only by the final class label. I intermediate valueIntermediate values tell us whether a finding arises from a "good"

geometric perturbation or a "bad" geometric perturbation in effective class priors. The evaluation records these quantities and associates them with minority recall, MCC, prior-response stability, tree disagreement, convergence over ensemble size and complexity, and computational complexity, as well as save the seed and partition identifiers necessary for repetition. This evidENTIALF structure is key for the claim boundary of the paper. It can say things about behavior in the named generator and doesn't turn simulated rows into clinical or financial or extent rows. The same distinction must be borne in mind when interpreting the runtime, which is a measurement from the execution environment itself, as opposed to a portable latency guarantee.

Reproducibility requires the source hash, seed ledger, profile table, split identifiers, fitted preprocessing state, algorithm parameters, result CSV files, and plot script. These items allow a numerical rerun to be conducted without turning controlled evidence into a benchmark claim. The design can be interpreted as restricted maximum likelihood estimation. The estimator takes as input a labeled partition that is imbalanced and outputs a classifier h and an audit state ω. The classifier is only meaningful as long as its class sensitive utility is above a threshold; the audit state is useful if a later reader can understand why some samples, splits, weights, or thresholds were accepted. PI-MHF encodes these conditions with finite candidate sets, fixed stopping rules, and numerical clipping. Corollaries provide boundedness or concentration under stated assumptions. They do not guarantee global optimality of the nonconvex forest, or superiority in general, or transfer to a completely unseen domain.

The main practical interpretation is conservative: run PI-MHF when the mecha-nism diagnostic points at the target problem, keep the documented fall back and evaluate the ultimate model with class-sensitive and probability based performance metrics both. This expression reveals how the research gap is connected to the proposed purpose. The gap is impurity reductions calculated from joint node counts contract as the minority prior decreases, hiding branch-conditionals that diverge. In particular, the aim is to introduce a Max-Hellinger node criterion, to show that it is invariant to changes of class-priors under fixed branch-conditionals, and to study a bounded bagged forest constituted by that criterion. A permissible test must separate the proposed statistic from class weighting, threshold tuning, and ensemble size. The experimental protocol provides these controls and the reports observed ordering. We then perform sensitivity analysis by varying one mechanism parameter at a time while holding the data generator, split, and metric definitions constant. This structure allows positive and negative findings to be interpreted in a bounded way and provides later studies using recorded data with a clear protocol to replicate.

VIII. CONCLUSION

We formulate PI-MHF for the goal of deriving a Max-Hellinger node rule, prove that the Max-Hellinger node rule is invariant to class-priors changes under fixed branch-conditionals, and study a finite bagged forest based on the Max-Hellinger node rule. There are three deterministic ordered algorithms that relate these measurable phenomena to conditional propositions. The related safeguard is of scientific interest as well as software interest. In the absence of a finite guard, an observation point with tiny local support, or a node containing very few rare-class samples, can overpower a ratio, distance, or posterior. Then a large numerical result may be interpreted as compelling evidence. PI-MHF avoids that failure by the bounds on the stage-specific stated in the method and restated in Table Safeguards. The bounds are the same in the proofs. The experiments measure their cost in terms of retention, calibration, disagreement, sensitivity, or fit time, so the safeguard isn’t an undocumented implementation detail.

Seeded the controlled study to show that PI-MHF is effective with limited applicability. It posts gains on several severe-profile diagnostics and losses on some other profile-metric pairs. The last claim is conditional and does not claim any registered benchmark or deployment superiority. The relationship between mechanism and evidence remains conditional. The results are governed by the proclaimed Gaussian-mixture structure, imbalance ratios, overlap settings, corruption model, and bounded learner class. These conditions are also helpful to analyze internals since the ground truth and corruption indicators are available. They are not a public benchmark rerun. The paper makes this distinction adjacent to each table and figure. Any subsequence application to recorded healthcare, fraud, intrusion, or edge data needs to refit preprocessing within the training partition, preserve entity or temporal groups, and report domain-specific costs prior to an operational conclusion be drawn from it.

Later evaluation should repeat the entire pipeline on public datasets with group-safe partitions, domain costs, larger repeated trials, calibration intervals, and external shift tests. The content at the present stage presents the equations, algorithms, profiles, and diagnostics needed for that follow-on study. From a decision-making perspective, false negative and false positive have different implications. One can buy somewhat higher rare-class

recall at completely unusable alert burden, and high precision may arise from classifying away most rare cases. We thus analyse recall alongside precision, MCC, average precision, balanced accuracy and calibration. PI-MHF is not selected by an "optimal" value but is assessed by the shape of this trade-off curve. The design moves selection of the threshold to after model fitting and limits it to calibration data, leaving the test partition out of all model and operating-point selection.

REFERENCES

[1] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263-1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[2] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," ACM Comput. Surv., vol. 49, no. 2, Art. no. 31, Aug. 2016, doi: 10.1145/2907070.

[3] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," Prog. Artif. Intell., vol. 5, pp. 221-232, 2016, doi: 10.1007/s13748-016-0094-0.

[4] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Syst. Appl., vol. 73, pp. 220-239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[5] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: Latest research, applications and future directions," Artif. Intell. Rev., vol. 57, 2024, doi: 10.1007/s10462-024-10759-6.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.

[7] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Advances in Intelligent Computing, LNCS 3644, 2005, pp. 878-887, doi: 10.1007/11538059_91.

[8] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE IJCNN, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[9] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," SIGKDD Explor., vol. 6, no. 1, pp. 20-29, 2004, doi: 10.1145/1007730.1007735.

[10] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in Proc. 14th Int. Conf. Mach. Learn., 1997, pp. 179-186.

[11] I. Tomek, "Two modifications of CNN," IEEE Trans. Syst., Man, Cybern., vol. SMC-6, no. 11, pp. 769-772, Nov. 1976, doi: 10.1109/TSMC.1976.4309452.

[12] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," IEEE Trans. Syst., Man, Cybern., vol. SMC-2, no. 3, pp. 408-421, Jul. 1972, doi: 10.1109/TSMC.1972.4309137.

[13] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," IEEE Trans. Syst., Man, Cybern. B, vol. 39, no. 2, pp. 539-550, Apr. 2009, doi: 10.1109/TSMCB.2008.2007853.

[14] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," IEEE Trans. Syst., Man, Cybern. A, vol. 40, no. 1, pp. 185-197, Jan. 2010, doi: 10.1109/TSMCA.2009.2029559.

[15] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," Pattern Recognit., vol. 40, no. 12, pp. 3358-3378, 2007, doi: 10.1016/j.patcog.2007.04.009.

[16] C. Elkan, "The foundations of cost-sensitive learning," in Proc. 17th Int. Joint Conf. Artif. Intell., 2001, pp. 973-978.

[17] L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, pp. 123-140, 1996, doi: 10.1007/BF00058655.

[18] L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[19] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119-139, 1997, doi: 10.1006/jcss.1997.1504.

[20] T. G. Dietterich, "Ensemble methods in machine learning," in Multiple Classifier Systems, LNCS 1857, 2000, pp. 1-15, doi: 10.1007/3-540-45014-9_1.

[21] S. Hido, H. Kashima, and Y. Takahashi, "Roughly balanced bagging for imbalanced data," Stat. Anal. Data Min., vol. 2, nos. 5-6, pp. 412-426, 2009, doi: 10.1002/sam.10061.

[22] Z. Liu, P. Cao, Z. Wang, J. Liu, and O. R. Zaiane, "Self-paced ensemble for highly imbalanced massive data classification," in Proc. IEEE 36th ICDE, 2020, pp. 841-852, doi: 10.1109/ICDE48307.2020.00077.

[23] F. Zhou, S. Gao, L. Ni, M. Pavlovski, Q. Dong, Z. Obradovic, and W. Qian, "Dynamic self-paced sampling ensemble for highly imbalanced and class-overlapped data classification," Data Min. Knowl. Discov., vol. 36, no. 5, pp. 1601-1622, 2022, doi: 10.1007/s10618-022-00838-z.
[24] D. A. Cieslak and N. V. Chawla, "Learning decision trees for unbalanced data," in Machine Learning and Knowledge Discovery in Databases, LNCS 5211, 2008, pp. 241-256, doi: 10.1007/978-3-540-87479-9_34.
[25] D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive," Data Min. Knowl. Discov., vol. 24, pp. 136-158, 2012, doi: 10.1007/s10618-011-0222-1.
[26] W. Liu, S. Chawla, D. A. Cieslak, and N. V. Chawla, "A robust decision tree algorithm for imbalanced data sets," in Proc. SIAM Int. Conf. Data Mining, 2010, pp. 766-777.
[27] R. Aler, O. Valls, and I. M. Galvan, "Study of Hellinger distance as a splitting metric for imbalanced data," Expert Syst. Appl., vol. 145, Art. no. 113090, 2020, doi: 10.1016/j.eswa.2019.113090.
[28] L. Baringhaus and C. Franz, "On a new multivariate two-sample test," J. Multivariate Anal., vol. 88, no. 1, pp. 190-206, 2004, doi: 10.1016/S0047-259X(03)00079-4.
[29] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, 1999, pp. 61-74.
[30] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers," in Proc. 18th Int. Conf. Mach. Learn., 2001, pp. 609-616.
[31] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proc. 8th ACM SIGKDD, 2002, pp. 694-699, doi: 10.1145/775047.775151.
[32] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Mach. Learn., 2005, pp. 625-632, doi: 10.1145/1102351.1102430.
[33] B. C. Wallace and I. J. Dahabreh, "Improving class probability estimates for imbalanced data," Knowl. Inf. Syst., vol. 41, pp. 33-52, 2014, doi: 10.1007/s10115-013-0670-6.
[34] M. Kull, T. M. Silva Filho, and P. Flach, "Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration," Electron. J. Stat., vol. 11, no. 2, pp. 5052-5080, 2017, doi: 10.1214/17-EJS1338SI.
[35] J. Tian, Y.-C. Liu, N. Glaser, Y.-C. Hsu, and Z. Kira, "Posterior re-calibration for imbalanced datasets," in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 8101-8113.
[36] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in Proc. Int. Conf. Learn. Representations, 2021.
[37] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in Proc. IEEE/CVF CVPR, 2019, pp. 9268-9277, doi: 10.1109/CVPR.2019.00949.
[38] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1567-1578.
[39] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.
[40] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, Art. no. 6, 2020, doi: 10.1186/s12864-019-6413-7.
[41] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861-874, 2006, doi: 10.1016/j.patrec.2005.10.010.
[42] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1-30, 2006.
[43] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.
[44] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," J. Mach. Learn. Res., vol. 18, no. 17, pp. 1-5, 2017.
[45] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in Proc. 20th Int. Conf. Pattern Recognit., 2010, pp. 3121-3124, doi: 10.1109/ICPR.2010.764.