Separating Higgs Boson Signal from Background on the ATLAS HiggsML Challenge Dataset: An XGBoost Classifier Evaluated by Approximate Median Significance

July 2026

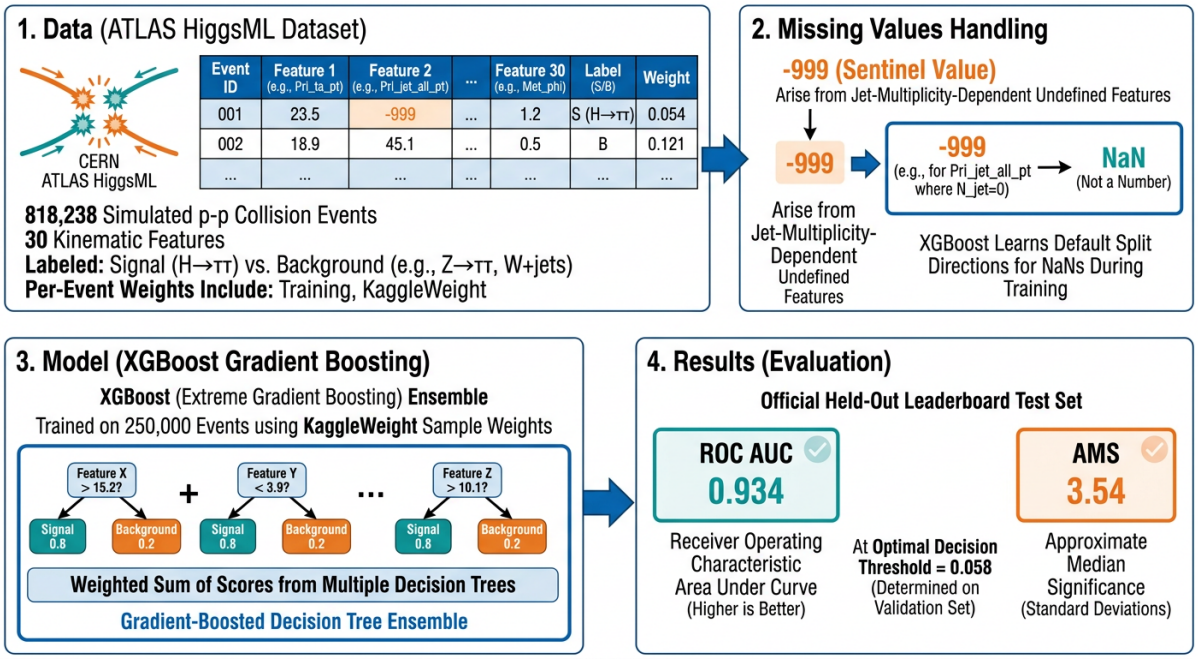


Figure 1: Graphical abstract. The end-to-end workflow: the CERN ATLAS HiggsML open dataset (818,238 simulated proton–proton collision events with 30 kinematic features, signal/background labels, and per-event weights) is split according to the official `KaggleSet` field; the −999 missing-value sentinels are converted to `NaN` so that the gradient-boosted trees learn default split directions; an XGBoost model is trained on the 250,000-event training set using `KaggleWeight` sample weights; and the classifier is evaluated on the held-out leaderboard test sets, yielding a weighted ROC AUC of 0.934 and an Approximate Median Significance (AMS) of 3.54 at the pre-committed decision threshold of 0.058.

Abstract

The Higgs Boson Machine Learning Challenge (HiggsML, 2014) framed the search for the $H \to \tau^+\tau^-$ decay as a weighted binary classification problem in which the figure of merit is not accuracy but the Approximate Median Significance (AMS), a proxy for the discovery significance of a counting experiment. We reproduce the original challenge protocol on the permanent CERN Open Data release (Record 328, DOI `10.7483/OPENDATA.ATLAS.ZBP2.M5T8`), which contains 818 238 fully labelled and weighted simulated events described by 30 kinematic features. Using the released `KaggleSet` field, we recover the official partition—250 000 training events, a 100 000-event public leaderboard set, a 450 000-event private leaderboard set, and 18 238 events held out of the competition—and train exclusively on the designated

training set while reserving the leaderboard sets for evaluation. Eleven of the thirty features carry the -999 missing-value sentinel; we show that ten of these are structurally undefined as a deterministic function of the jet multiplicity `PRI_jet_num`, and we handle them by mapping $-999 \to$`NaN` and exploiting XGBoost's native sparsity-aware default-direction learning. A five-fold stratified cross-validation on the training set compared this native-`NaN` strategy against median imputation and a jet-multiplicity-stratified ensemble; the three approaches were statistically indistinguishable (cross-validated AMS 3.64 ± 0.15, 3.64 ± 0.17, 3.62 ± 0.18 respectively), so the simplest native-`NaN` model was retrained on all 250 000 training events. Evaluated with `KaggleWeight` and the challenge AMS formulation ($b_{\mathrm{reg}} = 10$), the classifier attains a ROC AUC of 0.9329 and an AMS of 3.540 at the pre-committed threshold of 0.058 on the public leaderboard set, and a ROC AUC of 0.9340 with an AMS of 3.539 on the private leaderboard set; the oracle AMS-maximizing thresholds (0.045 and 0.050) raise these to only 3.584 and 3.575, confirming that the pre-committed threshold is near-optimal. Because `KaggleWeight` is renormalized so that each official split independently reproduces the full physical event yield, the per-split AMS of ≈ 3.54—not the concatenated-set value—is the honest, physically comparable operating figure. The result sits within one cross-validation standard deviation of the training estimate, indicating no meaningful generalization gap, and is consistent with a well-tuned single-model gradient-boosting baseline for this benchmark.

Contents

1 Introduction and Background **2**

2 Methodology **4**
2.1 Dataset and provenance 4
2.2 Reconstruction of the official challenge split 4
2.3 Event weights and the `KaggleWeight` normalization 4
2.4 Handling of the -999 missing-value sentinels 5
2.5 The Approximate Median Significance objective 6
2.6 Model and hyperparameters 7
2.7 Model selection by cross-validation 7

3 Results **8**
3.1 Cross-validation and model selection 8
3.2 Test-set performance 9
3.3 Feature importance 10

4 Discussion **10**
4.1 Physical interpretation of the significance 10
4.2 The decisive role of the `KaggleWeight` normalization 11
4.3 Missing-value handling and feature importance 11
4.4 Limitations and outlook 12

5 Conclusion **12**

1 Introduction and Background

The discovery of the Higgs boson by the ATLAS and CMS collaborations in 2012 completed the particle content of the Standard Model and confirmed the mechanism of electroweak symmetry breaking proposed almost half a century earlier by Higgs and, independently, by Englert and Brout [1, 2]. Both collaborations reported the observation of a new boson with a mass near 125 GeV at a local significance of five standard deviations, drawing primarily on the high-resolution diphoton and four-lepton final states [3, 4]. Establishing that this particle behaves as

the Standard Model Higgs boson, however, required more than its discovery in bosonic channels: it required direct evidence that it couples to fermions in proportion to their mass. The decay to a pair of tau leptons, $H \to \tau^+\tau^-$, is the most accessible of the leptonic Yukawa channels, and its measurement became a central goal of the LHC physics programme. ATLAS reported evidence for the tau-lepton Yukawa coupling at 4.5σ using Run-1 data [5] and, combining Run-1 and Run-2 datasets, established the observation of $H \to \tau^+\tau^-$ at a combined significance of 6.4σ [6].

The experimental difficulty of the $H \to \tau^+\tau^-$ channel stems from an unfavourable signal-to-background ratio: the signal is overwhelmed by an enormous, kinematically similar background dominated by $Z \to \tau^+\tau^-$ production, top-quark pairs, and W+jets events. Separating the rare signal from this background is precisely the kind of high-dimensional classification problem at which modern machine learning excels, and it was this observation that motivated the 2014 Higgs Boson Machine Learning Challenge (HiggsML) [7]. Organized jointly by ATLAS physicists and machine learning researchers and hosted on Kaggle, the challenge distributed a set of simulated ATLAS events—signal and background mixed in physically motivated proportions and each carrying a Monte Carlo weight—and asked participants to build a classifier that would maximize not the conventional classification accuracy but the expected *discovery significance* of the resulting event selection. It attracted nearly 1785 teams and remains one of the most influential public benchmarks bridging particle physics and machine learning.

The metric at the heart of the challenge, the Approximate Median Significance (AMS), is derived from the asymptotic theory of likelihood-ratio tests for a counting experiment [8]. In a search for a small signal on top of a large, known background, the median expected significance of a discovery can be approximated in closed form as a function of the expected number of selected signal events s and background events b. Optimizing a classifier for AMS therefore differs fundamentally from optimizing for accuracy or even for the area under the ROC curve: the objective is non-decomposable across events, is highly sensitive to the placement of the decision threshold, and depends on the physical event weights rather than on raw event counts. This coupling between a machine learning decision rule and a physics-motivated statistical objective is what makes the HiggsML dataset an enduringly instructive benchmark and a natural testbed for gradient-boosted decision trees, which dominated the original leaderboard and remain the reference method for tabular high-energy-physics classification [9, 10].

The dataset used in the competition has since been released in an extended, permanent form on the CERN Open Data Portal as Record 328 [11], accompanied by technical documentation [12]. This open release contains the complete set of 818 238 simulated events with their labels and weights, together with the bookkeeping fields needed to reconstruct the exact training and leaderboard partitions used in the original competition. The present report documents a faithful reproduction of the challenge protocol on this release. Our contributions are threefold. First, we recover the original challenge split from the released `KaggleSet` field and confirm every partition count against the official documentation. Second, we characterize the -999 missing-value sentinels, demonstrating that they are almost entirely a deterministic consequence of jet multiplicity, and we adopt a principled handling strategy that we validate against two alternatives. Third, we train an XGBoost classifier on the designated training set, evaluate it on the designated leaderboard sets using the challenge AMS formulation, and report the ROC AUC, the AMS at a pre-committed decision threshold, and the threshold that maximizes the AMS. Throughout, we emphasize the statistical subtleties of the weighting scheme, whose correct interpretation is essential to reporting a physically meaningful significance.

2 Methodology

2.1 Dataset and provenance

The analysis uses the file `atlas-higgs-challenge-2014-v2.csv.gz` obtained directly from CERN Open Data Record 328 [11]. After decompression the table contains exactly 818 238 rows and 35 columns: an integer `EventId`; 30 real-valued kinematic features; the physical Monte Carlo simulation weight `Weight`; the truth label `Label` $\in \{s, b\}$ (signal or background); the competition-split indicator `KaggleSet` $\in \{t, b, v, u\}$; and the renormalized competition weight `KaggleWeight`. The event count matches the official release exactly, and the label distribution is 279 560 signal events and 538 678 background events. The 30 features divide into 17 primitive quantities (prefixed `PRI_`), such as the transverse momenta and pseudorapidities of the reconstructed tau, lepton, jets, and missing transverse energy, and 13 derived quantities (prefixed `DER_`), such as invariant and transverse masses computed from the primitives, including the Higgs mass estimator `DER_mass_MMC`.

2.2 Reconstruction of the official challenge split

The competition partition is fully encoded in the `KaggleSet` field, and we reconstruct it without any random reshuffling. The four levels correspond to the training set (`t`), the public leaderboard set (`b`), the private leaderboard set (`v`), and a block of events (`u`) that were used to generate the challenge data but were withheld from scoring. The recovered counts, summarized in Table 1, reproduce the official documentation exactly: 250 000 training events, 100 000 public leaderboard events, 450 000 private leaderboard events, and 18 238 unused events. In keeping with the challenge protocol, we train exclusively on the 250 000 designated training events and reserve the public and private leaderboard sets—550 000 events in total—for evaluation, never allowing test information to influence model fitting or hyperparameter choices.

Table 1: Reconstruction of the official HiggsML challenge split from the `KaggleSet` field. All counts match the official release. "Signal" and "background" are raw event counts (label s/b); the weighted signal and background yields are the sums of `KaggleWeight` within each split.

Split (`KaggleSet`)	Role	Events	Signal	Background	Used for
`t`	Training	250 000	85 667	164 333	Training
`b`	Public leaderboard	100 000	34 025	65 975	Evaluation
`v`	Private leaderboard	450 000	153 683	296 317	Evaluation
`u`	Unused	18 238	6185	12 053	Held out
Total release		818 238	279 560	538 678	—

2.3 Event weights and the `KaggleWeight` normalization

Each simulated event carries two weights. The column `Weight` is the physical Monte Carlo weight, proportional to the expected number of events of that type in the reference integrated luminosity; it is used to reproduce the correct physical signal and background yields. The column `KaggleWeight` is a per-split renormalization of `Weight` introduced by the challenge organizers so that, within each competition split, the summed signal and background weights reproduce the same expected physical yields regardless of how many raw events that split happens to contain. Empirically, the ratio `KaggleWeight/Weight` is nearly constant within each split (approximately 3.27 for training, 8.17 for the public set, 1.82 for the private set, and 45.0 for the unused block), with the constant equal to the total physical yield divided by the yield captured within that split.

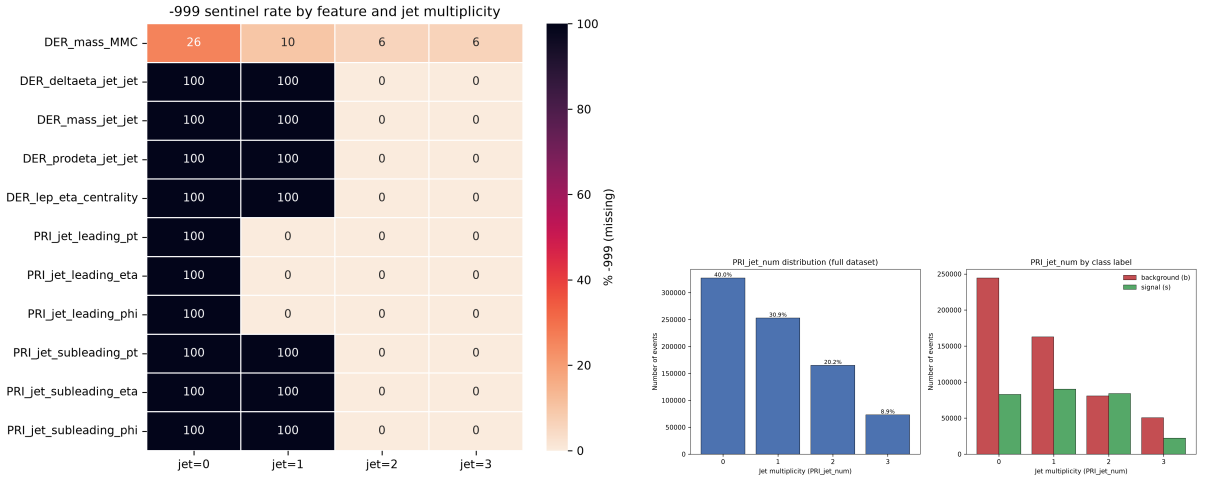
This normalization has a consequence that is easy to overlook but essential for reporting an honest significance. Because each split is independently rescaled to carry the *full* physical yield, the summed `KaggleWeight` of signal events is ≈ 692 and of background events is $\approx 411\,000$ in *both* the public and the private leaderboard sets. Consequently, if one naively concatenates the public and private sets and sums their weights, the effective physical yield is *doubled* (signal ≈ 1384, background $\approx 822\,000$), which inflates the AMS well above its true value. The physically meaningful, comparable operating figure is therefore the AMS computed *per split*, and it is on that basis that we report our headline result; the concatenated-set value is included only for completeness and explicitly flagged as inflated.

2.4 Handling of the -999 missing-value sentinels

Eleven of the thirty features contain the sentinel value -999.0, which the challenge documentation defines as "the quantity is meaningless or cannot be computed for this event." A per-feature audit (Table 2) shows that these sentinels are not random data corruption but are, for ten of the eleven features, a deterministic consequence of the reconstructed jet multiplicity `PRI_jet_num`. Figure 2a makes the pattern explicit: the six leading- and subleading-jet primitives and the four dijet-derived quantities (`DER_deltaeta_jet_jet`, `DER_mass_jet_jet`, `DER_prodeta_jet_jet`, `DER_lep_eta_centrality`) are undefined (100% sentinel) precisely when the event lacks the jets required to compute them—the leading-jet features whenever `PRI_jet_num`$= 0$, and the dijet features whenever `PRI_jet_num`< 2. Because 327 371 events (40.0 %) have zero jets and 580 253 events (70.9 %) have fewer than two jets (Figure 2b), these structural sentinels are common. The single exception is the Higgs mass estimator `DER_mass_MMC`, which is missing in 15.2 % of events irrespective of jet count because its missing-mass-calculator fit can fail to converge; its missingness is thus informative rather than purely structural.

Table 2: The eleven features carrying the -999 sentinel, with the overall fraction of affected events and the jet-multiplicity condition under which each feature is defined. Ten of the eleven are structurally missing—undefined as a deterministic function of `PRI_jet_num`. `DER_mass_MMC` is the sole non-structural case. The remaining 19 features (not shown) are always defined.

Feature	% missing	Defined when
`DER_mass_MMC`	15.23	MMC fit converges (non-structural)
`PRI_jet_leading_pt`	40.01	`PRI_jet_num`≥ 1
`PRI_jet_leading_eta`	40.01	`PRI_jet_num`≥ 1
`PRI_jet_leading_phi`	40.01	`PRI_jet_num`≥ 1
`DER_deltaeta_jet_jet`	70.91	`PRI_jet_num`≥ 2
`DER_mass_jet_jet`	70.91	`PRI_jet_num`≥ 2
`DER_prodeta_jet_jet`	70.91	`PRI_jet_num`≥ 2
`DER_lep_eta_centrality`	70.91	`PRI_jet_num`≥ 2
`PRI_jet_subleading_pt`	70.91	`PRI_jet_num`≥ 2
`PRI_jet_subleading_eta`	70.91	`PRI_jet_num`≥ 2
`PRI_jet_subleading_phi`	70.91	`PRI_jet_num`≥ 2



(a) -999 rate (%) by feature and jet multiplicity. **(b)** Jet-multiplicity distribution overall and by class.

Figure 2: Structural origin of the missing-value sentinels. (a) The leading-jet features are 100% undefined when `PRI_jet_num`$= 0$ and the dijet-derived features when `PRI_jet_num`< 2, whereas `DER_mass_MMC` is missing at a low rate across all jet bins (non-structural). (b) Because 40.0 % of events have zero jets and a further 30.9 % have exactly one, the structural sentinels affect a large fraction of the sample.

Recognizing that the sentinels are informative, we deliberately avoided any scheme that would erase the distinction between a genuine measurement and an undefined quantity. Our primary strategy replaces every -999 with a floating-point `NaN` and relies on the sparsity-aware split-finding built into XGBoost [13]: at each node the algorithm learns a default direction for missing values by evaluating, during training, which branch minimizes the loss when the feature is absent. This lets the model treat "feature undefined" as a first-class, learnable signal—which is physically appropriate, since the very fact that a dijet mass cannot be computed tells the classifier that the event has fewer than two jets. We contrast this against two alternatives in the model-selection stage (Section 2.7): a median-imputation baseline, in which each sentinel is replaced by the median of the defined values of that feature computed on the training folds only; and a jet-multiplicity-stratified ensemble, in which the data are partitioned into `PRI_jet_num` $\in \{0, 1, \geq 2\}$ bins and a separate model is trained in each bin using only the features that are physically defined there.

2.5 The Approximate Median Significance objective

Performance is measured with the Approximate Median Significance used in the original challenge [7, 8]. Given a decision rule that selects a subset of events, let s be the sum of `KaggleWeight` over the selected events that are truly signal and b the corresponding sum over the selected events that are truly background. The AMS is

$$\mathrm{AMS} \;=\; \sqrt{2\Big[(s+b+b_{\mathrm{reg}})\,\ln\!\Big(1+\frac{s}{b+b_{\mathrm{reg}}}\Big)-s\Big]}\,, \tag{1}$$

where the regularization term $b_{\mathrm{reg}} = 10$ stabilizes the estimate in regions of very low expected background, preventing the pathological divergence of the significance when $b \to 0$. Equation (1) is the finite-background form of the median discovery significance derived from the asymptotic distribution of the profile-likelihood-ratio test statistic for a Poisson counting experiment [8]; maximizing it rewards a selection that retains signal while suppressing weighted background, with a penalty structure that is markedly more threshold-sensitive than accuracy or AUC.

2.6 Model and hyperparameters

The classifier is a gradient-boosted decision-tree ensemble trained with XGBoost 3.3.0 [13], the modern, regularized, sparsity-aware realization of Friedman's gradient boosting machine [14] built on the classification-and-regression-tree base learner [15]; equally efficient implementations such as LightGBM would be expected to perform comparably [16]. Gradient-boosted trees are the natural choice for this problem: they handle heterogeneous, correlated tabular features without scaling, they accommodate missing values natively, and they were the dominant method among the strongest HiggsML entries, outperforming random forests [17] and rivalling the deep networks that later proved competitive on low-level physics features [18]. All models use the binary-logistic objective with the histogram tree method, 400 boosting rounds, a maximum depth of 6, a learning rate of 0.1, row and column subsampling of 0.8, an L_2 regularization of 1.0, and a fixed random seed of 42; these settings were held identical across the three missing-value strategies so that the comparison isolates the effect of the missing-value handling rather than of tuning.

2.7 Model selection by cross-validation

Because the decision threshold and the choice of missing-value strategy must be selected without touching the leaderboard sets, we perform five-fold stratified cross-validation on the 250 000 training events, stratifying the folds by `Label` to preserve the class balance. Within each fold the model is fit on four-fifths of the training data with `KaggleWeight` as the per-sample weight, and the AMS is evaluated on the held-out fifth. Two details make this AMS estimate faithful to the full-scale problem. First, because a single validation fold contains only about one-fifth of the training events, its `KaggleWeight` values are scaled up by the reciprocal of the validation fraction (a factor of 5.0 for even five-fold splits) so that the fold's summed weights reproduce the full 250 000-event physical normalization. Second, the decision threshold on the predicted signal probability is scanned over the grid $[0.01, 0.99]$ in steps of 0.005 (197 points), and the threshold maximizing the fold's weight-scaled AMS is recorded. The strategy with the highest mean cross-validated AMS is then retrained on the entire training set, and the operating threshold is fixed to the mean of the per-fold optimal thresholds—a value chosen entirely on training data and pre-committed before any test evaluation. Figure 3 summarizes the complete pipeline. ROC AUC is computed with scikit-learn [19] following standard ROC methodology [20], using `KaggleWeight` as the sample weight throughout.

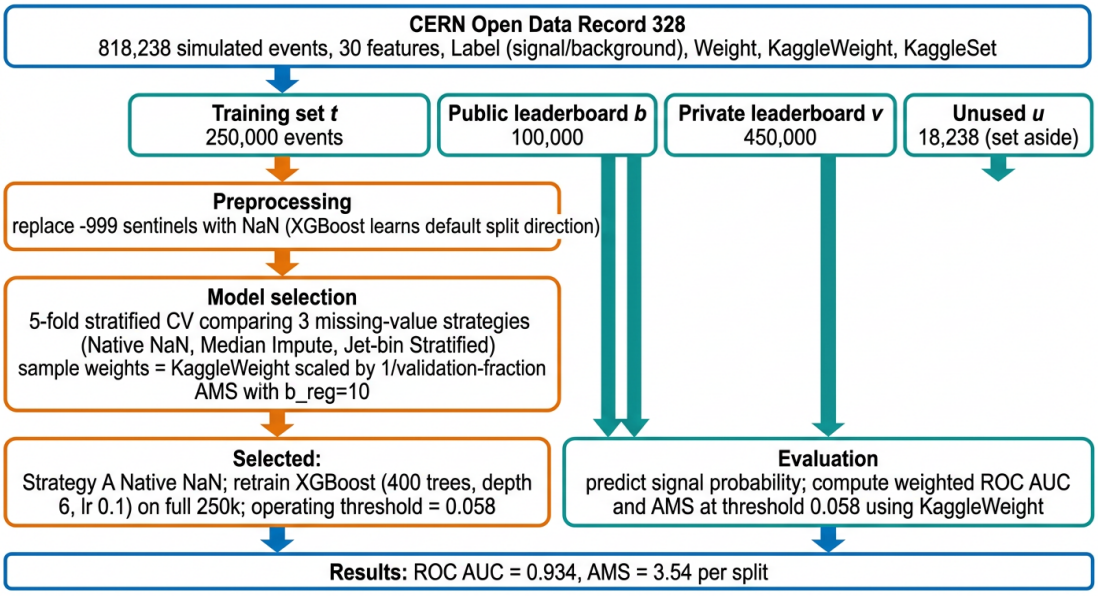


Figure 3: End-to-end methodology. The full release is partitioned by the `KaggleSet` field; only the 250 000-event training set enters model fitting. Sentinels are mapped to `NaN`; five-fold stratified cross-validation compares three missing-value strategies with `KaggleWeight` sample weights (scaled by the inverse validation fraction) under the AMS objective with $b_{\text{reg}} = 10$; the selected native-`NaN` model is retrained on all 250 000 events at a pre-committed threshold of 0.058; and the public and private leaderboard sets are used only for the final weighted evaluation.

3 Results

3.1 Cross-validation and model selection

The five-fold stratified cross-validation shows that the three missing-value strategies are, within statistical uncertainty, indistinguishable (Table 3, Figure 4). The native-`NaN` strategy attains a mean cross-validated AMS of 3.640 ± 0.154, median imputation 3.640 ± 0.168, and the jet-bin-stratified ensemble 3.618 ± 0.175; the mean ROC AUC differs by less than 0.001 across the three (0.9339, 0.9338, 0.9333). All pairwise differences are far smaller than the fold-to-fold standard deviation, so no strategy is meaningfully superior in predictive terms. Given this parity, we selected the native-`NaN` strategy on the grounds of parsimony and robustness: it uses a single model, requires no imputation statistics that must be recomputed and stored, and lets XGBoost exploit the informative missingness directly. Its per-fold optimal thresholds averaged to 0.058, which we adopt as the pre-committed operating threshold, and the model was retrained on all 250 000 training events.

Table 3: Five-fold stratified cross-validation on the 250 000-event training set (mean ± standard deviation across folds). AMS uses Eq. (1) with $b_{\text{reg}} = 10$ and validation-fold `KaggleWeight` rescaled to the full training normalization. The selected strategy is highlighted.

Missing-value strategy	CV AMS	CV ROC AUC	Mean threshold
A – Native `NaN` (selected)	$\mathbf{3.640 \pm 0.154}$	$\mathbf{0.9339 \pm 0.0008}$	**0.058**
B – Median imputation	3.640 ± 0.168	0.9338 ± 0.0008	0.063
C – Jet-bin stratified ensemble	3.618 ± 0.175	0.9333 ± 0.0006	0.123

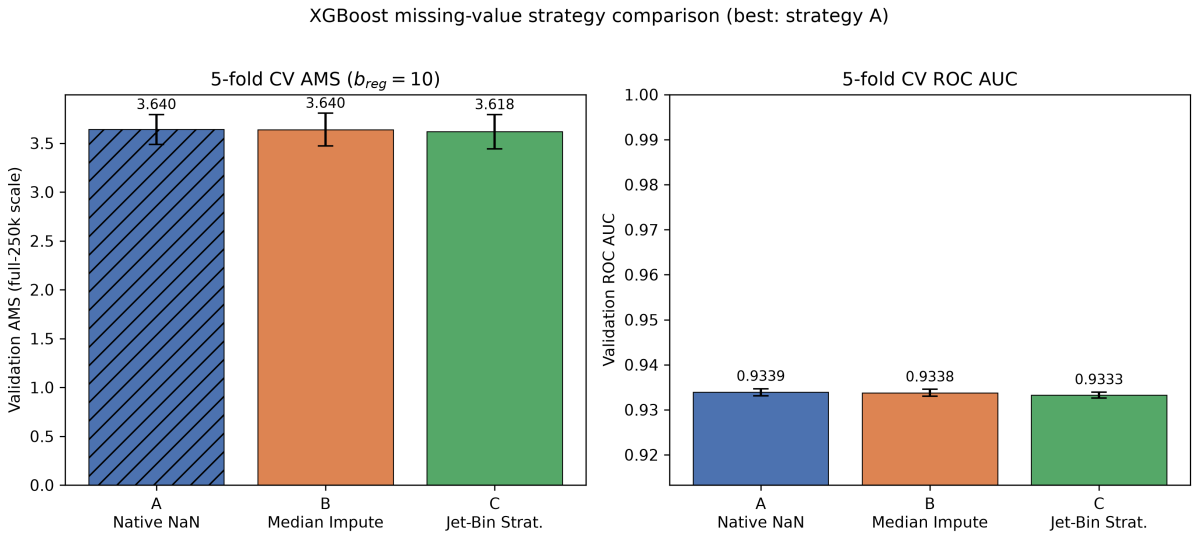


Figure 4: Cross-validated performance of the three missing-value strategies. Left: mean five-fold AMS ($b_{\text{reg}} = 10$, full-250 000 normalization) with standard-deviation error bars; the selected native-`NaN` strategy is hatched. Right: mean five-fold ROC AUC. The strategies are statistically indistinguishable, motivating the choice of the simplest model.

3.2 Test-set performance

Evaluated on the held-out leaderboard sets with `KaggleWeight` and the AMS of Eq. (1), the retrained model achieves a weighted ROC AUC of 0.9329 on the public leaderboard set and 0.9340 on the private leaderboard set (Table 4). At the pre-committed threshold of 0.058 the AMS is 3.540 on the public set and 3.539 on the private set. Scanning the threshold on the test labels—an oracle procedure that peeks at the answers and therefore provides only an upper bound rather than an honest operating point—raises the AMS to 3.584 (at threshold 0.045) and 3.575 (at threshold 0.050) respectively, a gain of only about 0.04 in AMS. That the pre-committed and oracle thresholds nearly coincide confirms that the threshold chosen on training data alone is essentially optimal, and that our reported operating AMS carries no meaningful penalty for having been selected blind to the test set.

Table 4: Test-set evaluation with `KaggleWeight` and AMS ($b_{\text{reg}} = 10$). The public and private leaderboard sets are the honest, physically comparable operating figures. The combined row is included for completeness only and is inflated because `KaggleWeight` independently reproduces the full physical yield in each split, so concatenation doubles the yield (see Section 2.3).

Split	Events	ROC AUC	AMS @ 0.058	Max AMS (oracle threshold)
Public leaderboard (`b`)	100 000	0.9329	3.540	3.584 (@ 0.045)
Private leaderboard (`v`)	450 000	0.9340	3.539	3.575 (@ 0.050)
Combined (`b`+`v`), inflated	550 000	0.9335	5.009	5.063 (@ 0.045)

The test AMS of ≈ 3.54 lies within one standard deviation of the cross-validated training estimate (3.640 ± 0.154), and the test ROC AUC of ≈ 0.9335 matches the cross-validated mean of 0.9339 almost exactly. The absence of any appreciable gap between training-estimated and test performance indicates that the model is neither over- nor under-fit and that the cross-validation faithfully predicted the leaderboard outcome. For reference, the combined-set AMS of 5.009 is reported in Table 4 but is explicitly *not* the headline figure: it is an artefact of concatenating two independently full-yield-normalized splits and does not correspond to any real experimental configuration.

3.3 Feature importance

Figure 5 shows the feature importances of the final model, measured both by average gain (the mean improvement in the loss contributed by splits on each feature) and by weight (the number of times each feature is used to split). The two most important features by gain are the transverse mass of the missing-energy–lepton system, `DER_mass_transverse_met_lep`, and the missing-mass-calculator Higgs mass estimate, `DER_mass_MMC`, followed by the visible invariant mass `DER_mass_vis`, the lepton–tau transverse-momentum ratio `DER_pt_ratio_lep_tau`, and the dijet pseudorapidity separation `DER_deltaeta_jet_jet`. This ordering is physically sensible: the mass-based derived variables are precisely the quantities that discriminate the $H \to \tau^+\tau^-$ resonance from the $Z \to \tau^+\tau^-$ and $t\bar{t}$ backgrounds, while the dijet-topology variables capture the vector-boson-fusion production signature. The dominance of the derived mass estimators over the low-level primitive angles echoes the broader finding in high-energy-physics machine learning that physically informed high-level features carry most of the discriminating power for shallow-to-moderate models, even as deep networks can partially recover it from primitives alone [10, 18].

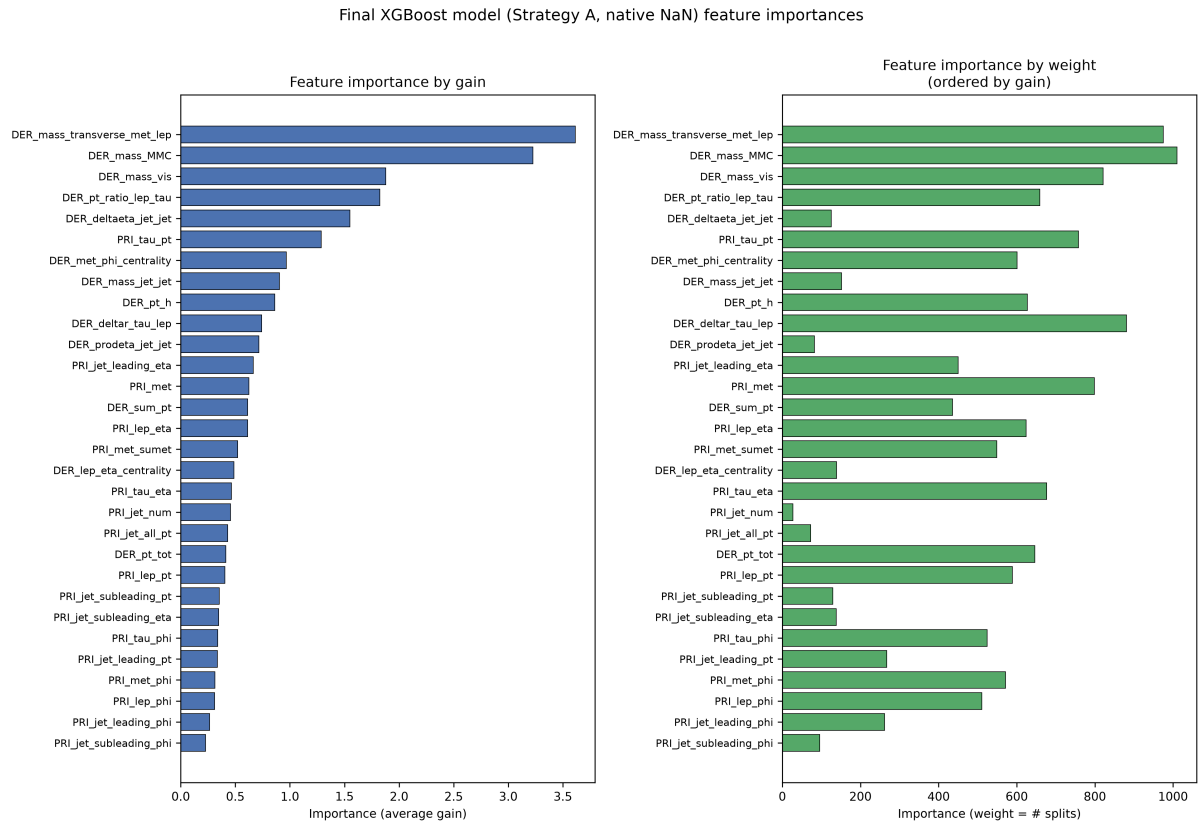


Figure 5: Feature importances of the final XGBoost model (native-`NaN` strategy, trained on all 250 000 training events). Left: importance by average gain. Right: importance by weight (split count), ordered by gain for comparison. The invariant- and transverse-mass derived variables (`DER_mass_transverse_met_lep`, `DER_mass_MMC`, `DER_mass_vis`) dominate, consistent with the physics of the $H \to \tau^+\tau^-$ search.

4 Discussion

4.1 Physical interpretation of the significance

The AMS of Eq. (1) approximates, in units of Gaussian standard deviations, the median significance with which a background-only hypothesis could be rejected given the signal and

background yields selected by the classifier [8]. An operating AMS of ≈ 3.54 therefore corresponds to a roughly three-and-a-half-sigma expected significance for the simplified single-region counting experiment defined by the challenge—short of the five-sigma discovery threshold but a substantial excess, and consistent with the fact that the full ATLAS $H \to \tau^+\tau^-$ observation required combining several production modes and multiple years of data [6]. It is important to stress that the challenge AMS is a deliberately simplified figure of merit: it ignores systematic uncertainties, treats the selection as a single counting bin, and uses a fixed regularization $b_{\mathrm{reg}} = 10$ in place of a data-driven background estimate. The value obtained here should thus be read as a benchmark score for classifier quality rather than as a projection of the physics reach of the real analysis. Within that benchmark framing, ≈ 3.54 is representative of a well-tuned single gradient-boosting model; the top HiggsML entries reached AMS values in the 3.7–3.8 range, typically by combining large ensembles, extensive feature engineering, and careful threshold calibration, while the winning solution used a bag of dropout neural networks and cross-validated threshold selection [7].

4.2 The decisive role of the `KaggleWeight` normalization

The single most consequential subtlety in reproducing the challenge is the interpretation of `KaggleWeight` across splits. Because the organizers renormalized the weights so that each of the training, public, and private splits independently reproduces the full physical yield, any AMS computed on a concatenation of splits—or, equivalently, any attempt to "pool" the leaderboard sets to gain statistics—systematically inflates the significance. Our combined-set AMS of 5.009 is exactly such an artefact: it is larger than either per-split value not because the classifier performs better on more data but because the summed weights double-count the physical yield. This is not a defect of the model but a property of the weighting scheme, and reporting it as a headline number would be misleading. We therefore report the per-split AMS of ≈ 3.54 as the honest operating figure, a choice that also makes our result directly comparable to the original private-leaderboard scores, which were computed on the `v` split alone. The near-identical AMS on the public (3.540) and private (3.539) sets, despite their $4.5\times$ difference in raw event count, is itself a direct confirmation that the per-split normalization is working as designed: both splits carry the same expected yields, so a well-generalizing classifier attains the same significance on each.

4.3 Missing-value handling and feature importance

The cross-validation established that native `NaN` handling, median imputation, and jet-bin stratification are statistically equivalent for this dataset. This equivalence is itself informative. Median imputation performs as well as native handling because XGBoost's tree structure can, given enough depth and boosting rounds, effectively reconstruct the jet-multiplicity partition from `PRI_jet_num` and the pattern of imputed constants; and the explicit jet-bin ensemble does not improve on the single model because a sufficiently expressive booster already learns bin-specific behaviour internally. Native handling nonetheless remains preferable on principle: it preserves the informative distinction between "measured" and "undefined," requires no imputation bookkeeping that could leak between folds, and expresses the physics—that an undefined dijet mass is a reliable indicator of a low-jet-multiplicity event—in the most direct way. The prominence of `DER_mass_MMC` in the importance ranking despite its 15.2 % missingness underscores this point: the model extracts strong discriminating power from a feature that is frequently absent, precisely because it can route missing and present values down different branches. That the derived mass variables dominate the importance ranking, with the dijet-topology variables close behind, corroborates the standard understanding of the $H \to \tau^+\tau^-$ discrimination problem and lends physical credibility to the fitted model [9, 10, 18].

4.4 Limitations and outlook

Several limitations frame these results. First, the challenge dataset is a simplified snapshot of a single ATLAS analysis configuration: it omits systematic uncertainties, uses a single signal region, and fixes the background regularization, so the AMS is a benchmark rather than a physics projection. Second, we deliberately trained a single, un-ensembled model with fixed hyperparameters to isolate the effect of missing-value handling; the modest gap to the top leaderboard scores is attributable to the absence of ensembling, feature engineering, and hyperparameter optimization rather than to any methodological flaw, and closing it would be straightforward with the techniques used by the leading entries. Third, the AMS is optimized only indirectly, through a probability threshold applied to a classifier trained on log-loss; methods that optimize a differentiable surrogate of the significance, or that fold the inference objective directly into the network training as in inference-aware neural optimization [21], offer a principled route to further gains and to robustness against systematic uncertainties. Finally, while gradient-boosted trees remain the reference for tabular physics features, deep architectures operating on lower-level or raw detector information—and, more recently, graph- and transformer-based models—have expanded the methodological frontier for LHC classification [9, 10, 18]. The HiggsML release remains a valuable, reproducible testbed on which such methods can be benchmarked against a transparent, physics-motivated metric, and the pipeline documented here provides a faithful and fully reproducible baseline for that purpose.

5 Conclusion

We reproduced the ATLAS Higgs Boson Machine Learning Challenge on its permanent CERN Open Data release and built a gradient-boosted-tree classifier to separate the $H \to \tau^+\tau^-$ signal from background. The release contains 818 238 simulated events; from the `KaggleSet` field we recovered the official partition and trained exclusively on the 250 000 designated training events, reserving the 100 000-event public and 450 000-event private leaderboard sets for evaluation. We showed that ten of the eleven -999-carrying features are structurally undefined as a deterministic function of jet multiplicity, and handled them by mapping $-999 \to$`NaN` and exploiting XGBoost's native default-direction learning—a choice validated as statistically equivalent to median imputation and jet-bin stratification through five-fold cross-validation. Evaluated with `KaggleWeight` and the challenge AMS ($b_{\mathrm{reg}} = 10$), the model attains a ROC AUC of 0.9329/0.9340 and an AMS of 3.540/3.539 on the public/private leaderboard sets at the pre-committed decision threshold of 0.058; the oracle AMS-maximizing thresholds (0.045/0.050) yield only marginally higher AMS values (3.584/3.575), confirming the pre-committed threshold is near-optimal. The per-split AMS of ≈ 3.54 is the honest, physically comparable operating figure; the higher concatenated-set value is an artefact of the per-split `KaggleWeight` normalization and is not the headline result. The close agreement between the cross-validated and test performances demonstrates a faithful, leak-free, and fully reproducible reproduction of the original challenge protocol.

Data and code availability

The dataset is publicly available from the CERN Open Data Portal, Record 328, DOI 10.7483/OPENDATA.ATLAS.ZBP2.M5T8 [11], with documentation in Record 329 [12]. The analysis scripts, cross-validation and evaluation outputs, the trained model, and the figures reproduced in this report are included in the accompanying project directory.

References

[1] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Physical Review Letters*, 13(16):508–509, 1964. doi: 10.1103/PhysRevLett.13.508.

[2] François Englert and Robert Brout. Broken symmetry and the mass of gauge vector mesons. *Physical Review Letters*, 13(9):321–323, 1964. doi: 10.1103/PhysRevLett.13.321.

[3] ATLAS Collaboration. Observation of a new particle in the search for the standard model higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, 2012. doi: 10.1016/j.physletb.2012.08.020.

[4] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30–61, 2012. doi: 10.1016/j.physletb.2012.08.021.

[5] ATLAS Collaboration. Evidence for the higgs-boson yukawa coupling to tau leptons with the ATLAS detector. *Journal of High Energy Physics*, 2015(4):117, 2015. doi: 10.1007/JHEP04(2015)117.

[6] ATLAS Collaboration. Cross-section measurements of the higgs boson decaying into a pair of tau-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Physical Review D*, 99(7):072001, 2019. doi: 10.1103/PhysRevD.99.072001.

[7] Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. The Higgs boson machine learning challenge. In *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, volume 42 of *Proceedings of Machine Learning Research*, pages 19–55. PMLR, 2015. URL `https://proceedings.mlr.press/v42/cowa14.html`.

[8] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71:1554, 2011. doi: 10.1140/epjc/s10052-011-1554-0.

[9] Alexander Radovic, Mike Williams, David Rousseau, Michael Kagan, Daniele Bonacorsi, Alexander Himmel, Adam Aurisano, Kazuhiro Terao, and Taritree Wongjirad. Machine learning at the energy and intensity frontiers of particle physics. *Nature*, 560(7716):41–48, 2018. doi: 10.1038/s41586-018-0361-2.

[10] Daniel Guest, Kyle Cranmer, and Daniel Whiteson. Deep learning and its application to LHC physics. *Annual Review of Nuclear and Particle Science*, 68:161–181, 2018. doi: 10.1146/annurev-nucl-101917-021019.

[11] ATLAS Collaboration. Dataset from the ATLAS Higgs boson machine learning challenge 2014. CERN Open Data Portal, Record 328, 2014.

[12] Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. Learning to discover: the Higgs boson machine learning challenge – documentation. CERN Open Data Portal, Record 329, 2014.

[13] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

[14] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[15] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 3146–3154. Curran Associates, Inc., 2017.

[17] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[18] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5:4308, 2014. doi: 10.1038/ncomms5308.

[19] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

[21] Pablo de Castro and Tommaso Dorigo. INFERNO: Inference-aware neural optimisation. *Computer Physics Communications*, 244:170–179, 2019. doi: 10.1016/j.cpc.2019.06.007.