Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning-Based Higgs Boson Event Classification Using CERN ATLAS Data

Tvisha Dhuper
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee https://orcid.org/0000-0001-8829-9980

Method Article

Keywords: Higgs boson, H → ττ, collider physics, machine learning, gradient boosting, ATLAS Open Data, Approximate Median Significance, bootstrap inference, SHAP

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11695868/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Machine Learning-Based Higgs Boson Event Classification Using CERN ATLAS Open Data

Tvisha Dhuper, and Mayank Goswami [1, 2, #]
[1]Divyadrishti Imaging Laboratory, Department of Physics, IIT Roorkee
[2]*Mehta Family School of Data Science and Artificial Intelligence*, IIT Roorkee,
Roorkee, Uttarakhand-247667, India.

[#]mayank.goswami@ph.iitr.ac.in orcid{0000-0001-8829-9980}

Abstract

The decay of the Higgs boson into a tau-lepton pair ($H \rightarrow \tau\tau$) is challenging to isolate because each tau decay produces at least one neutrino, which prevents complete event reconstruction, while several background processes yield similar visible final states. This study evaluates machine-learning classifiers for simulated $H \rightarrow \tau\tau$ events using the ATLAS Higgs Machine Learning Challenge dataset released through CERN Open Data. The analysis uses 818,238 events, 30 original input features, and the official training, validation, and test partitions. Performance is assessed using weighted ROC-AUC and the Approximate Median Significance (AMS), with the operating threshold selected exclusively on the validation set. Nonlinear classifiers outperform logistic regression: random forests, multilayer perceptrons, and gradient-boosted trees achieve test AUC values close to 0.93, compared with 0.8997 for logistic regression. The engineered-feature gradient-boosting model reaches a test AMS of 3.579, with a 95% bootstrap interval of approximately 3.52–3.66. Paired bootstrap comparisons indicate that the observed gains from hyperparameter tuning and feature engineering are not statistically distinguishable from zero, whereas per-jet specialised models perform significantly worse. A four-model ensemble obtains an AMS of 3.648 in a reduced-computation experiment; this result is promising but requires confirmation with full training configurations. SHAP analysis identifies the estimated parent mass as the dominant predictor, with its contribution largest in the vicinity of the Higgs-boson mass. The results support a reproducible and statistically cautious baseline rather than a claim of state-of-the-art performance.

Keywords: Higgs boson; $H \rightarrow \tau\tau$; collider physics; machine learning; gradient boosting; ATLAS Open Data; Approximate Median Significance; bootstrap inference; SHAP

1 Introduction

The ATLAS and CMS collaborations announced the discovery of a new boson with a mass near 125 GeV in 2012 [1, 2], subsequently established as the Higgs boson. A central objective of the ensuing programme has been to test whether the particle couples to fermions with strengths consistent with the Standard Model. The decay $H \rightarrow \tau\tau$ provides an important probe of the Higgs coupling to charged leptons. Evidence for this decay was reported by both collaborations [3, 4], followed by its observation [5, 6].

Isolating $H \rightarrow \tau\tau$ events remains challenging. Each tau lepton decays into visible particles and at least one neutrino; the neutrinos escape direct detection, making complete reconstruction of the parent resonance difficult. Moreover, the irreducible $Z \rightarrow \tau\tau$ background produces a closely related final state. Additional backgrounds arise from processes such as top-quark pair production and W + jets. Their kinematic distributions overlap with those of the signal, so discrimination depends on combining several observables rather than relying on a single variable. Machine-learning classifiers offer a way to exploit nonlinear relationships among these observables. Importantly, the classifier threshold determines the balance between signal retention and background rejection and therefore affects the expected discovery sensitivity.

This work develops and evaluates a reproducible classification pipeline using the public ATLAS challenge data. Beyond comparing predictive scores, it examines whether apparent improvements are robust to paired statistical testing and whether the learned decision patterns are physically interpretable.

The contributions are: (i) a pipeline following the official data partition; (ii) a comparison of linear, bagged, neural-network, and boosted-tree classifiers; (iii) paired bootstrap comparisons of model performance; and (iv) a SHAP-based examination of the relationship between model predictions and event kinematics. The study is intended as a transparent benchmark and does not claim to surpass the strongest entries in the original challenge.

2 Problem Objectives and Need

The task is to classify each simulated collision event as signal ($H \rightarrow \tau\tau$) or background using its measured and derived kinematic features. Each event is assigned a weight representing its expected contribution to the collision sample. Consequently, an unweighted classification error does not necessarily correspond to an equal change in the expected signal or background yields. The practical objective is therefore to identify a score threshold that provides strong signal-background separation while optimising a significance-oriented metric, rather than maximising raw accuracy.

Objectives.

1. Build a leakage-free, reproducible classification pipeline on the official split.
2. Compare linear, bagged, neural and boosted models with a physics-motivated metric (AMS).
3. Test whether tuning, engineered features, per-jet modelling and ensembling help, using paired statistical tests.
4. Explain the best model with SHAP and check that it uses sensible physics.
5. Provide code and a deployable model so that others can reproduce and extend the results.

Why it is needed.

- Higgs signal events are rare relative to background processes. Consequently, the quality of the selected event region is more relevant to a search than overall classification accuracy.
- Performance differences among strong nonlinear classifiers are often small. Confidence intervals and paired comparisons are therefore essential to distinguish reproducible improvements from fluctuations caused by the finite test sample.
- A classifier intended for use in a physics analysis should be scrutinised for physically plausible behaviour. Understanding the influence of reconstructed masses, transverse momenta, angular variables, and event topology can support validation and reveal possible modelling artefacts.
- Open, documented baselines are valuable for teaching and for benchmarking new methods.

3 Theory

3.1 Weighted classification and ROC-AUC

Let x denote the feature vector for an event, $y = 1$ indicate signal and $y = 0$ indicate background, and w denote the event weight. A classifier returns a score f(x). The weighted receiver operating characteristic (ROC) curve displays the weighted signal efficiency against the weighted background acceptance as the score threshold is varied. The corresponding weighted area under the curve (AUC) can be interpreted as the probability that a randomly selected signal event receives a higher score than a randomly selected background event, with the sampling probabilities determined by the event weights. AUC quantifies ranking performance across thresholds; it does not, by itself, identify the operating point that maximises discovery sensitivity.

3.2 Discovery significance and the AMS

For a selection with expected signal s and expected background b, the asymptotic median significance for discovery is [15]

$$Z = \sqrt{2\left((s+b)\,ln\,ln\,\left(1+\frac{s}{b}\right)-s\right)} \qquad (1)$$

The challenge metric adds a regularisation term $b_r = 10$ to the background, which stabilises the metric when b is small [9]:

$$AMS = \sqrt{2\left((s+b+b_r)\,ln\,ln\,\left(1+\frac{s}{b+b_r}\right)-s\right)} \qquad (2)$$

Here, s and b are the sums of the event weights for selected signal and background events, respectively. The challenge definition includes a background regularisation term (b_r = 10) to avoid unstable behaviour when the selected background yield is small. In the regime s ≪ b, the metric behaves approximately as s/√b, illustrating the trade-off it imposes: retaining signal is beneficial, but rejecting background can be especially valuable. AMS is therefore threshold-dependent, unlike AUC. In this study, the threshold is optimised on validation data and then held fixed for the test evaluation.

3.3 Gradient boosting

Gradient boosting [10] builds an additive model

$$F_m(x) = F_{m-1}(x) + \nu h_m(x), \qquad (3)$$

where each small decision tree h_m is fitted to the negative gradient of the loss at the current model and ν is a learning rate. For classification the loss is the logistic loss, and F is a log-odds. Histogram-based implementations (scikit-learn), XGBoost [11] and LightGBM [12] discretise features for speed and handle missing values natively, which suits this data set.

3.4 Other learners

Logistic regression models the log-odds as a linear function of standardised features and is the linear baseline. A random forest averages many decorrelated deep trees. A multilayer perceptron is a feed-forward neural network trained by stochastic gradient descent.

3.5 Probability calibration

Platt scaling [16] maps a raw score f to a probability by logistic regression on held-out data,

$$p = \frac{1}{1+exp(Af+B)}. \qquad (4)$$

It matters when scores from different models must be compared against one common cut, as in our per-jet experiment.

3.6 Bootstrap and paired bootstrap

The bootstrap estimates sampling variability by repeatedly resampling test events with replacement and recalculating the metric. For a paired comparison, the same resampled event indices are used for both models. This preserves event-level correspondence and reduces the influence of fluctuations shared by the two predictions. We regard a difference as statistically distinguishable from zero when its 95% bootstrap interval excludes zero. This criterion is an interval-based assessment, not a substitute for a pre-specified multiple-comparison procedure; the number of bootstrap replicates also limits the precision of the interval endpoints.

3.7 SHAP explanations

The Shapley value of feature i assigns it the average marginal change in the model output when it is added to all possible subsets S of the other features:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)]. \quad (5)$$

SHAP values use Shapley-value concepts to decompose a model prediction into feature-level contributions relative to a reference output [13, 14]. TreeSHAP computes these attributions efficiently for tree ensembles. In this analysis, contributions are expressed in log-odds units: a positive value moves the prediction towards the signal class, while a negative value moves it towards background. Mean absolute SHAP values summarise global attribution magnitude, whereas beeswarm and dependence plots show the distribution and direction of contributions across events. These quantities explain the fitted model; they do not establish causal relationships between a feature and the underlying physics process.

4 Previous Work

The discovery of the Higgs boson by ATLAS and CMS [1, 2] initiated detailed studies of its properties. Evidence for the H → ττ decay was reported by ATLAS [3] and CMS [4], followed by measurements establishing the decay channel [5, 6]. Such analyses exploit multivariate information, including boosted decision trees, to distinguish signal from backgrounds with similar final states.

Machine learning has become an established component of collider-data analysis. Baldi, Sadowski, and Whiteson demonstrated the potential of deep neural networks for particle-physics classification tasks [7], while Guest, Cranmer, and Whiteson reviewed developments and applications of deep learning at the LHC [8]. These studies motivate comparisons between neural and tree-based approaches, although performance is strongly dependent on the dataset, event representation, and evaluation objective.

The 2014 ATLAS Higgs Machine Learning Challenge [9] provided simulated H → ττ events and popularised the use of the Approximate Median Significance as an optimisation metric. The leading challenge submissions reported AMS values around 3.8, often using ensembles of neural networks and boosted decision trees [9]. The dataset was subsequently made available through the CERN Open Data Portal [18], enabling reproducible educational and methodological studies.

Boosted trees and explainability. Gradient boosting [10] and its fast implementations XGBoost [11] and LightGBM [12] are standard for tabular data. SHAP [13] and TreeSHAP [14] provide consistent attributions for such models.

The present study is positioned as a reproducible baseline rather than a competition entry. Its emphasis is on consistent data partitioning, validation-based threshold selection, uncertainty estimation, paired model comparisons, and interpretation of the fitted classifier. The cited works provide the relevant background, but this section is not intended to constitute an exhaustive systematic review.

5 Method

5.1 Data

The dataset comprises 818,238 simulated events from the ATLAS Higgs Machine Learning Challenge [18]. Each event contains 17 primary (PRI) and 13 derived (DER) kinematic features, together with an event weight, class label, and split identifier. The sentinel value −999 denotes quantities that are undefined for a given event (for example, jet-related quantities when no jet is reconstructed); these values were treated as missing rather than as physical measurements. The official partition was retained: 250,000 training events (code t), 100,000 validation events (b), and 450,000 test events (v). A further 18,238 events (code u) were not used. Although approximately one-third of events are signal by event

count, the signal constitutes a much smaller fraction of the total event weight. This distinction motivates the use of weighted evaluation metrics.

5.2 Preprocessing and features

Undefined values were converted to missing entries. Tree-based algorithms were allowed to handle missing values natively, whereas models requiring complete numerical inputs used median imputation. Missingness indicators were added for two features. Log-transformed copies were generated for selected heavy-tailed mass and momentum variables. Raw azimuthal-angle (φ) variables were excluded because their absolute values depend on the arbitrary choice of coordinate origin; angular differences are more directly interpretable and invariant under a common azimuthal rotation. In the extended experiment, nine additional variables were constructed: six absolute angular separations (tau–lepton, tau–missing momentum, and lepton–missing momentum, together with the corresponding separations involving the leading jet) and three ratios (tau/lepton transverse momentum, missing transverse momentum/tau transverse momentum, and estimated mass/visible mass). The resulting extended representation contains 49 inputs, including the log-transformed copies. Class-balanced sample weights were used during training. Feature engineering was treated as an empirical hypothesis to be tested, not as an assumed improvement.

5.3 Models

- Logistic regression (standardised, median-imputed).
- Random forest, 300 trees, minimum leaf size 5.
- Multilayer perceptron, hidden layers 128 and 64, early stopping.
- Histogram gradient boosting (scikit-learn): learning rate 0.05, 31 leaves, up to 600 iterations with early stopping; a random search of 10 trials over learning rate, leaves, minimum leaf size and L2 regularisation, scored by validation AMS.
- Per-jet models: one gradient-boosting model each for events with 0, 1 and 2 or more jets, with and without Platt calibration (fitted on a held-out 20% of the training set).
- XGBoost, LightGBM, and a PyTorch network (layers 256, 128, 64, dropout 0.2, weighted cross-entropy), each with early stopping on a 10% hold-out of the training set.
- Averaged ensembles (arithmetic mean of predicted probabilities).

5.4 Evaluation protocol

All models were trained using the training partition. For each model, the score threshold that maximised AMS was selected using the validation partition and then applied without modification to the test partition. Thus, reported test AMS values represent performance at a validation-selected operating point. The test partition was not used for threshold selection or model choice. This separation reduces direct test-set optimisation, although repeated experimentation on the validation set can still introduce selection effects.

5.5 Statistical comparison

Uncertainty intervals were estimated from 300 bootstrap resamples of the test events. Paired comparisons used identical resampled events for both models, and the reported interval refers to the distribution of the difference in AMS. Five-fold cross-validation on the training partition was used as a supplementary check of weighted-AUC stability. The bootstrap analysis quantifies event-sampling variability under the supplied sample and does not incorporate detector, theory, or other systematic uncertainties.

5.6 Explanation

TreeSHAP values were calculated for a gradient-boosted model trained with the engineered features, using 3,000 randomly selected validation events. The resulting global and dependence plots are used to examine which inputs influence the fitted model and whether the direction of their contributions is qualitatively compatible with the expected signal kinematics.

5.7 Software and reproducibility

Python with scikit-learn, XGBoost, LightGBM, PyTorch and SHAP. Every script has an offline smoke-test mode and fixed random seeds. See Data and Code Availability.

5.8 Hypothesis-driven questions

The experiments are guided by four testable hypotheses. H1: nonlinear classifiers outperform a linear decision boundary because signal-background separation depends on interactions among reconstructed masses, transverse momenta, and angular observables. H2: a reconstructed parent-mass estimator provides the strongest individual discrimination, with signal-like contributions concentrated around the Higgs-boson mass scale. H3: adding physically motivated angular and momentum-ratio features improves the classifier's access to event geometry, although any improvement in the significance-oriented metric may be smaller than the gain in global ranking performance. H4: training separate classifiers by jet multiplicity may capture category-specific structure, but the reduction in sample size per category may offset this benefit; alternatively, averaging diverse classifiers may reduce variance and improve stability. These hypotheses are assessed using the held-out test results, paired bootstrap intervals, and SHAP plots. The analyses provide varying levels of support and should not be interpreted as causal tests.

6 Results

6.1 Baseline models

Table 1: Baseline models.

Model	AUC (validation)	AUC (test)	AMS (test, validation cut)
Logistic regression	0.8983	0.8997	2.63
Random forest	0.9275	0.9288	3.44
Neural network (MLP)	0.9285	0.9291	3.45
Gradient boosting	0.9282	0.9295	3.57

Across the baseline comparison, all nonlinear methods outperform logistic regression in both weighted AUC and AMS. The increase in test AUC from 0.8997 for logistic regression to 0.9295 for gradient boosting indicates that the class boundary is not adequately represented by a simple linear combination of the input variables. Random forest and the MLP also reach AUC values near 0.929, suggesting that the improvement is not unique to one nonlinear architecture. The relatively narrow performance range among these three models indicates that, once nonlinear interactions are captured, differences between model families are modest on this dataset. The close agreement between validation and test AUC provides a useful consistency check, although it does not by itself establish generalisation to detector data or a different simulation campaign.

6.2 Tuning, per-jet models and ensembling

Table 2: Tuning, per-jet models and ensembling.

Model	AUC (test)	AMS (test, validation cut)	95% CI
Baseline gradient boosting	0.9295	3.567	3.50 to 3.64
Tuned gradient boosting	0.9307	3.588	3.51 to 3.68
Per-jet gradient boosting	0.9248	3.407	3.34 to 3.48
Neural network (MLP)	0.9291	3.453	3.39 to 3.52
Ensemble (tuned, per-jet, MLP)	0.9304	3.540	3.48 to 3.60

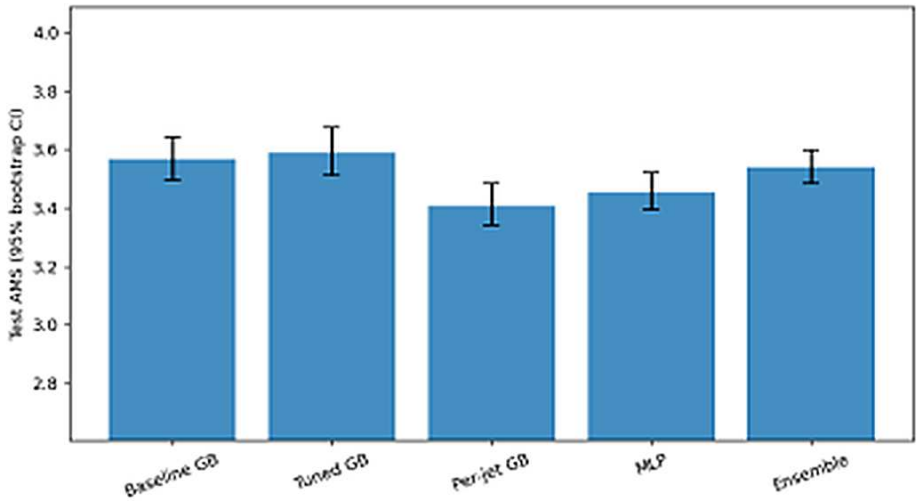


Figure 1. Test-set AMS for the baseline, tuned, per-jet, MLP, and ensemble configurations. Error bars denote 95% bootstrap intervals. The tuned model's interval overlaps that of the baseline, so the small numerical increase should not be interpreted as a demonstrated improvement. The per-jet configuration has a visibly lower AMS, consistent with the paired comparison reported below.

6.3 Cross-validation, engineered features and calibrated per-jet models

The five-fold training-set cross-validation gives a weighted AUC of 0.9299 ± 0.0009, with fold values between 0.9283 and 0.9311. This low fold-to-fold dispersion is consistent with the validation and test estimates and suggests that the ranking performance is not strongly dependent on a particular training fold. It should nevertheless be interpreted as an internal stability check, because the folds are drawn from the same simulated sample and do not probe distribution shifts.

Table 3: Engineered features and calibrated per-jet models.

Model	AUC (test)	AMS	Change vs baseline (paired 95% CI)	Significant
Baseline gradient boosting	0.9295	3.567	—	—
+ engineered features	0.9313	3.579	+0.011 (−0.025 to +0.048)	no
Per-jet, calibrated	0.9235	3.363	−0.209 (−0.260 to −0.153)	yes
Average: baseline + per-jet calibrated	0.9282	3.520	−0.048 (−0.091 to −0.009)	yes

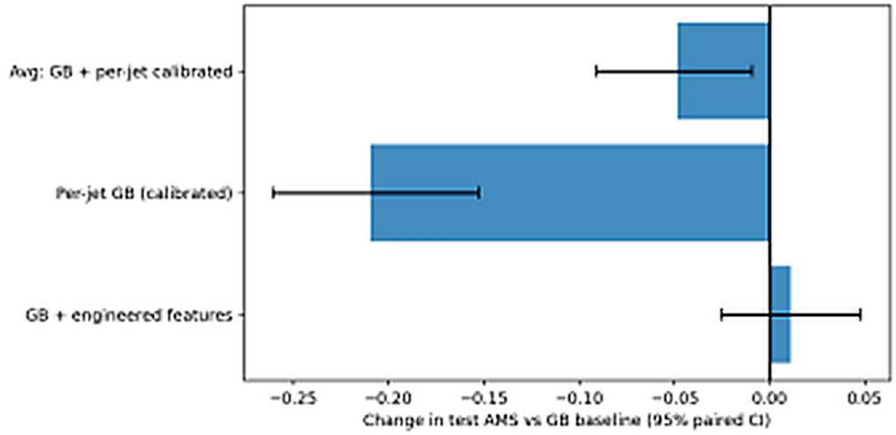


Figure 2. Paired change in test AMS relative to the baseline gradient-boosting model. The horizontal error bars show 95% paired-bootstrap intervals. The engineered-feature interval crosses zero, indicating that its observed gain is inconclusive. In contrast, the calibrated per-jet model and its average with the baseline have intervals below zero, supporting a reduction in AMS relative to the reference.

Adding the engineered variables increases test AUC from 0.9295 to 0.9313, but the paired AMS difference is +0.011 with a 95% interval from −0.025 to +0.048. The interval includes zero; hence, the data do not establish a significant improvement in the significance-oriented metric. This distinction is important: a modest gain in global ranking quality need not translate into a better operating point at the validation-selected threshold. Calibration also fails to recover the performance lost by the per-jet models. The negative paired interval suggests that dividing the sample into jet categories reduces the effective training information available to each classifier or otherwise makes the specialised decision boundaries less effective. These are plausible explanations, not mechanisms directly established by the present experiment.

6.4 Performance by jet multiplicity

Table 4: Performance by jet multiplicity.

Jets	Test events	Signal fraction	AUC (ensemble)
0	179,993	0.252	0.9389
1	138,925	0.357	0.9117
2 or more	131,082	0.448	0.9235

The event-count signal fraction is highest in the two-or-more-jet category (0.448), followed by the one-jet (0.357) and zero-jet (0.252) categories. However, the ensemble AUC is highest for zero-jet events (0.9389), while one-jet events are the most difficult to rank (0.9117). This contrast shows that a larger signal fraction does not necessarily imply easier classification: separability depends on the overlap of the signal and background feature distributions, not simply on class prevalence. The comparatively strong zero-jet performance may reflect cleaner or more characteristic kinematics in this category, whereas the one-jet topology may contain a more heterogeneous mixture of production and background processes. The latter interpretation is a hypothesis and would require category-specific distributions and process breakdowns to verify.

6.5 Boosting libraries, neural network and a larger ensemble

The following comparison uses the engineered feature set and the same official data partition. It is based on a reduced-computation configuration, with at most 150 trees per boosted model and at most four epochs for the neural network. These results are therefore exploratory and should not be treated as

equivalent to fully trained model results. In particular, early stopping and the restricted training budgets may affect the relative ranking of the implementations.

Table 5: Boosting libraries, neural network and ensemble (reduced-size configuration).

Model	AUC (test)	AMS	Change vs reference (paired 95% CI)	Significant
scikit-learn GB (reference)	0.9309	3.547	—	—
XGBoost	0.9314	3.592	+0.044 (+0.011 to +0.079)	yes
LightGBM	0.9317	3.578	+0.030 (−0.008 to +0.068)	no
PyTorch MLP	0.9322	3.557	+0.008 (−0.049 to +0.063)	no
Ensemble of the four	0.9328	3.648	+0.102 (+0.066 to +0.138)	yes

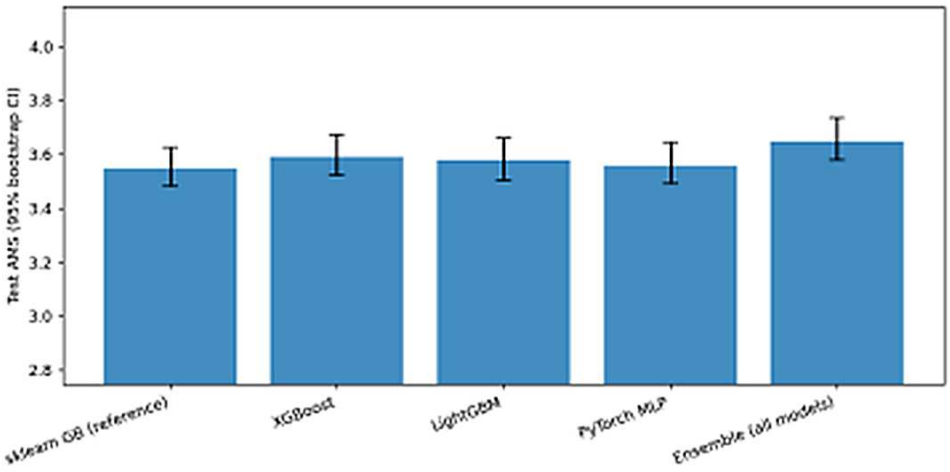


Figure 3. Test-set AMS for the scikit-learn gradient-boosting reference, XGBoost, LightGBM, PyTorch MLP, and their four-model probability-averaging ensemble. Error bars represent 95% bootstrap intervals. The reduced-size training configuration is an important qualification when interpreting these comparisons.

Among the individual implementations, the AMS values are close, and the paired intervals for LightGBM and the PyTorch MLP include zero relative to the reference. XGBoost shows a positive paired difference in this reduced-size experiment (+0.044; 95% interval +0.011 to +0.079). The four-model ensemble reaches the largest AMS (3.648), with a paired gain of +0.102 (95% interval +0.066 to +0.138). A plausible explanation is that averaging predictions from learners with different inductive biases reduces model-specific variance and stabilises event ranking near the selection threshold. However, the ensemble result comes from a restricted training budget; it should be regarded as a promising hypothesis for follow-up full-budget evaluation rather than as a definitive improvement over the final single-model result. This ensemble also differs from the earlier ensemble, which included a weaker per-jet classifier.

6.6 Final single-model selection

The selected deployable single model is gradient boosting with engineered features. Its operating threshold (0.8947) was fixed by maximising AMS on the validation partition. Applied to the 450,000 test events, it selects 70,708 events (15.7%), with an event-count purity of 92.3% and a signal efficiency of 42.5%. The weighted signal and background yields are 237.7 and 4,322.3, respectively, giving an

AMS of 3.579. These quantities describe the challenge sample and its event weights; they should not be interpreted as measured yields or discovery significance in collision data.

Table 6: Final single-model selection on the test set.

Quantity	Value
Events selected	70,708 of 450,000 (15.7%)
Purity by event count	92.3%
Signal efficiency	42.5%
Weighted signal *s*	237.7
Weighted background *b*	4,322.3
AMS	3.579

The selected sample is highly enriched in signal by event count (92.3%), while the signal accounts for only about 5.2% of the selected total event weight. This apparent difference arises because background events carry substantially larger aggregate weights in the supplied sample. It illustrates why event-count purity alone is an inadequate measure of search performance and why the weighted yields enter the AMS calculation. The independently recomputed AMS of 3.5792 agrees with the evaluation script, providing an internal consistency check on the saved predictions and threshold implementation.

6.7 What the model learned (SHAP)

Table 7: Top ten features by mean absolute SHAP value.

Rank	Feature	Mean absolute SHAP
1	DER_mass_MMC	0.748
2	DER_mass_transverse_met_lep	0.383
3	DER_mass_vis	0.308
4	PRI_tau_pt	0.235
5	dphi_tau_lep (engineered)	0.215
6	DER_mass_MMC_log	0.201
7	DER_met_phi_centrality	0.123
8	PRI_lep_eta	0.117
9	PRI_jet_leading_eta	0.116
10	dphi_lep_met (engineered)	0.097

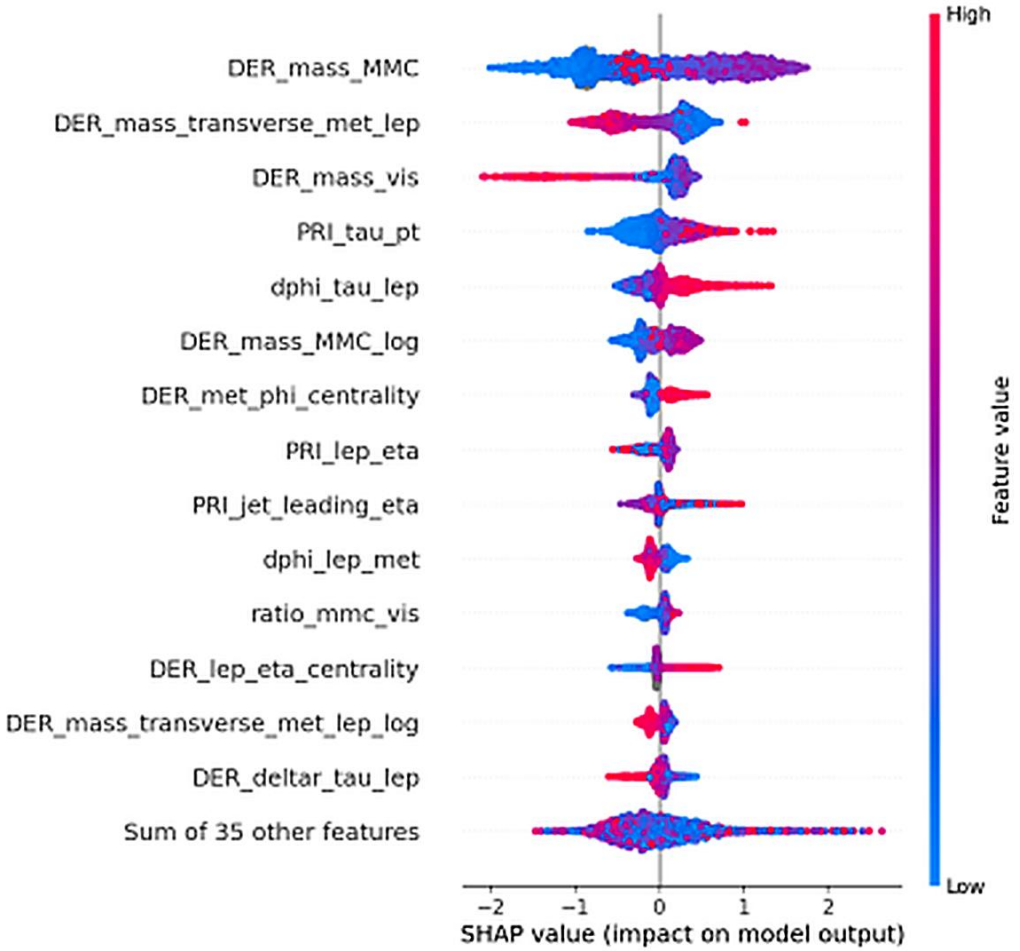


Figure 4. SHAP beeswarm summary for the ten most influential individual features and the aggregate contribution of the remaining features. Each point represents one event; horizontal position gives the SHAP contribution to the model output (positive towards signal, negative towards background), while colour encodes the feature value. The plot displays both the relative global importance and the event-to-event variation in feature effects.

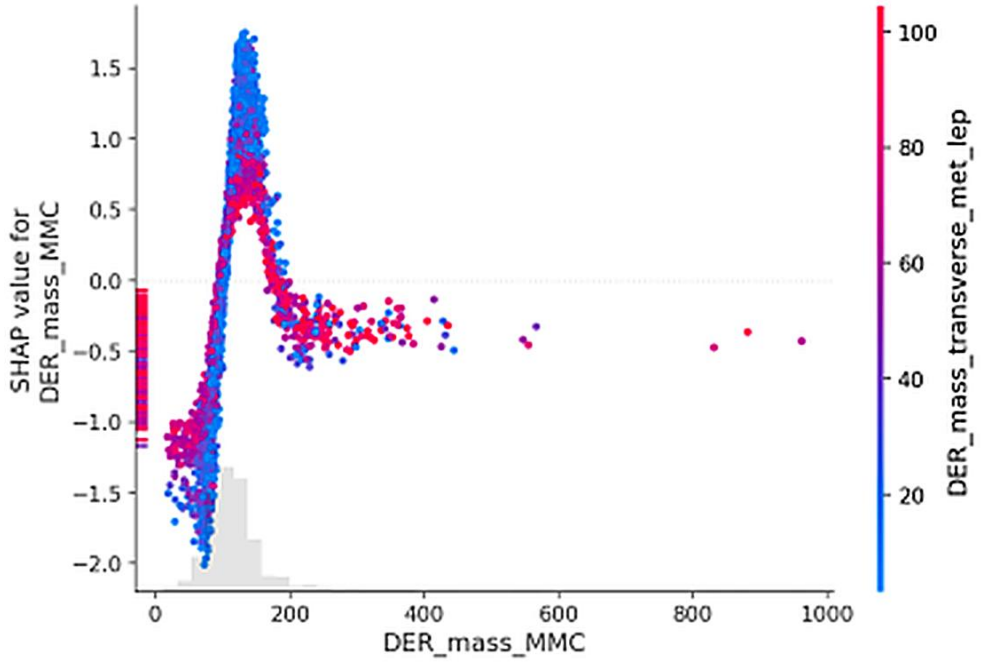


Figure 5. SHAP dependence plot for DER_mass_MMC. The horizontal axis is the estimated parent mass, the vertical axis is its SHAP contribution to the model output, and point colour represents DER_mass_transverse_met_lep. The non-monotonic response indicates that the model associates a restricted mass region with stronger signal-like predictions.

The estimated parent mass (DER_mass_MMC) is the most influential feature, with a mean absolute SHAP value of 0.748, substantially larger than that of the next-ranked variable, DER_mass_transverse_met_lep (0.383). In Figure 5, the contribution from DER_mass_MMC is strongly negative at low reconstructed mass, rises to a positive maximum of approximately +1.7 log-odds units in the region of roughly 110–150 GeV, and decreases again at higher mass, reaching around −0.4 above 200 GeV. This non-monotonic pattern is physically plausible: a classifier searching for a resonance near 125 GeV should favour events whose reconstructed parent-mass estimate is compatible with that scale, while assigning less signal-like support to substantially lower or higher values. The broad peak should not be interpreted as a direct measurement of the Higgs mass or as a detector-resolution estimate; it is a learned contribution conditional on the other inputs and on the training sample. The log-transformed copy of the same mass variable also appears among the leading features. Because these two inputs are correlated, their SHAP attributions divide the available explanatory contribution, and their individual importance values should not be interpreted independently.

6.8 Synthesis of the evidence and working hypotheses

Taken together, the results provide differentiated support for the hypotheses stated in Section 5.8. H1 is supported by the consistent advantage of nonlinear learners over logistic regression, although the present comparison does not identify which specific feature interactions account for the gain. H2 is supported by the dominant global SHAP importance of DER_mass_MMC and by the local maximum in its contribution near the Higgs mass. H3 receives partial support: engineered angular and ratio variables appear among the influential features and slightly increase AUC, but the paired AMS interval includes zero. H4 is not supported for the per-jet strategy in its present form, which reduces AMS; the diversity-ensemble result is encouraging but remains provisional because it was obtained under reduced training budgets. These findings suggest that the most useful next step is not simply to increase model complexity, but to test whether the observed patterns persist under independent samples, full training budgets, and realistic detector variations.

Figure 4 indicates that the visible mass and transverse-mass-related variables contribute strongly to event classification, while tau transverse momentum and angular separations also carry useful

information. The tau–lepton angular difference ranks fifth by mean absolute SHAP value, and the lepton–missing-momentum angular difference appears among the top ten. These results are consistent with the hypothesis that signal-background discrimination depends not only on a resonance-mass proxy but also on the geometry of the visible decay products and missing transverse momentum. The SHAP distributions are not uniformly monotonic: the same feature value can contribute differently in different regions of the feature space, reflecting nonlinear interactions. The importance of angular variables is also compatible with the small improvement in AUC after feature engineering, although the paired AMS interval does not demonstrate a statistically reliable gain in the selected operating point.

The event-level explanations provide illustrative, rather than population-level, evidence. For the signal-like example (classifier score 0.995), an estimated mass of 133.6 GeV contributes +1.45 log-odds units, with additional positive contributions from a large dijet mass and dijet pseudorapidity separation. Such a topology can be associated with vector-boson-fusion production, but a single SHAP explanation cannot establish the production mechanism. For the background-like example (score 0.005), a visible mass of 187.6 GeV contributes −0.87 log-odds units, while an undefined mass estimate contributes −0.5. The latter illustrates that missingness itself can be informative to a model, but it also warrants scrutiny: the model may be exploiting event-topology or reconstruction patterns encoded by missing values. Dedicated checks would be needed to exclude unwanted dependence on detector or simulation-specific missingness patterns.

7 Conclusion

This study establishes a reproducible machine-learning baseline for separating simulated $H \rightarrow \tau\tau$ events from background in the ATLAS Higgs Machine Learning Challenge dataset. The nonlinear classifiers achieve test weighted-AUC values close to 0.93 and AMS values around 3.6, clearly outperforming the logistic-regression baseline. The improvement indicates that the event-level discrimination problem contains nonlinear relationships and feature interactions that are not captured adequately by a linear decision function. Among the fully described single-model experiments, gradient boosting with engineered features gives a test AMS of 3.579 at a threshold selected on validation data. The observed numerical gains from hyperparameter tuning and feature engineering are not statistically conclusive under the paired-bootstrap analysis, whereas the per-jet specialised approach performs significantly worse. Thus, greater model complexity or finer event partitioning does not automatically improve the physics-oriented objective.

The reduced-computation comparison suggests that probability averaging across different model families may improve robustness: its AMS of 3.648 is accompanied by a positive paired-bootstrap interval. Because this result was obtained with restricted training budgets, it remains provisional pending full-budget replication. The SHAP analysis provides a complementary physics-oriented result. The estimated parent mass dominates the global attribution, and its contribution is largest in a mass region surrounding the known Higgs scale. Angular and momentum-related variables provide additional discrimination, supporting a multivariate interpretation in which resonance compatibility and event topology jointly shape the classifier response. These attributions demonstrate consistency between the learned decision function and selected kinematic expectations, but they do not prove that the model has learned a unique or causal representation of Higgs production.

The principal contribution is therefore not a new performance record, but a transparent workflow that combines weighted evaluation, validation-only threshold selection, uncertainty estimation, paired comparisons, and model interpretation. The results provide a useful starting point for further studies using full training budgets, independent simulation samples, and detector-level validation.

8 Limitations

- The study uses simulated challenge data rather than collision data. Consequently, the reported performance may not capture detector conditions, reconstruction imperfections, pile-up, or differences between simulation and real observations.
- The AMS values are calculated using the challenge's event weights and regularisation convention. They are useful for comparing classifiers within this dataset but are not a measurement of the statistical significance of a Higgs signal in an experimental analysis. A realistic search would require a likelihood-based treatment of nuisance parameters and systematic uncertainties.
- Most model comparisons use one fixed official training/validation/test partition. Five-fold cross-validation was used only to assess training-set AUC stability; it was not a complete nested cross-validation procedure for all model choices and thresholds. Repeated experimentation may also lead to some adaptation to the validation sample.
- Hyperparameter optimisation was limited to a random search of ten trials for the principal gradient-boosting model. The XGBoost, LightGBM, PyTorch, and four-model ensemble comparisons were conducted with reduced training budgets and therefore require full-length replication before firm conclusions can be drawn about their relative performance.

9 Future directions

Future work should first repeat the reduced-budget library and ensemble comparisons with fully converged training and multiple random seeds. Robustness should then be evaluated using independent simulated samples, alternative generator and detector configurations, pile-up variations, and systematic shifts in reconstructed observables. Category-wise studies should separate jet multiplicity from production mode and background composition, while missing-value ablation tests should determine whether undefined quantities provide physically meaningful information or encode reconstruction-specific artefacts. Finally, a physics-analysis deployment would require calibration studies, nuisance-parameter-aware statistical inference, and validation on collision data. These steps are necessary before interpreting the classifier as an analysis-ready Higgs-selection tool.

- The performance differences among the strongest individual classifiers are small. The available bootstrap analysis does not justify a definitive ranking of all implementations, and the reported intervals quantify event-resampling uncertainty rather than all sources of experimental uncertainty.
- The per-jet models use smaller category-specific training samples. Reduced sample size is a plausible explanation for their poorer performance, but this mechanism was not isolated experimentally from differences in category composition, feature distributions, or model optimisation.
- SHAP values explain the fitted model's predictions and do not establish causal relationships. Correlated variables, including the reconstructed mass and its logarithmic copy, can share attribution. The analysis also uses a finite subset of validation events and does not quantify explanation stability across independently trained models.
- The study does not report a dedicated robustness analysis under alternative detector conditions, systematic feature variations, class-prior shifts, or simulation-to-data domain adaptation. Nor does it provide a detailed breakdown of performance by individual background process or Higgs production mode. The literature review is limited to the cited works and is not systematic.

Data and Code Availability

The data are available from the CERN Open Data Portal (record 328) [18]. The code is organised as: higgs_classification.py (baselines), advanced_experiments.py (tuning, per-jet, ensemble),

extended_analysis.py (cross-validation, engineered features, paired tests, saved model), boosting_and_nn.py (XGBoost, LightGBM, PyTorch), explain_shap.py, predict.py, app.py (web demo) and higgs_walkthrough.ipynb.

Video Resources/Commentary

https://youtu.be/8DeiFE7zW1o

References

[1] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Phys. Lett. B* 716 (2012) 1–29.

[2]CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys. Lett. B* 716 (2012) 30–61.

[3] ATLAS Collaboration, "Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector," *JHEP* 04 (2015) 117.

[4] CMS Collaboration, "Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons," *JHEP* 05 (2014) 104.

[5]CMS Collaboration, "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector," *Phys. Lett. B* 779 (2018) 283–316.

[6] ATLAS Collaboration, "Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton–proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector," *Phys. Rev. D* 99 (2019) 072001.

[7] P. Baldi, P. Sadowski, D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature Communications* 5 (2014) 4308.

[8] D. Guest, K. Cranmer, D. Whiteson, "Deep learning and its application to LHC physics," *Annu. Rev. Nucl. Part. Sci.* 68 (2018) 161–181.

[9] C. Adam-Bourdarios et al., "The Higgs boson machine learning challenge," *Proceedings of Machine Learning Research* 42 (2015) 19–55.

[10] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics* 29 (2001) 1189–1232.

[11] T. Chen, C. Guestrin, "XGBoost: a scalable tree boosting system," *Proc. KDD* (2016) 785–794.

[12] G. Ke et al., "LightGBM: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems* 30 (2017).

[13] S. M. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems* 30 (2017).

[14] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence* 2 (2020) 56–67.

[15] G. Cowan, K. Cranmer, E. Gross, O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *Eur. Phys. J. C* 71 (2011) 1554.

[16] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers* (1999).

[17] B. Efron, "Bootstrap methods: another look at the jackknife," *Annals of Statistics* 7 (1979) 1–26.

[18] ATLAS Collaboration, "ATLAS Higgs Boson Machine Learning Challenge 2014 data set," CERN Open Data Portal, record 328, https://opendata.cern.ch/record/328.