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.

Neural Network-Based Classification of Higgs Boson Signal and Background Events in the H→τ τ Channel Using ATLAS Open Data

DHRUV KUMAR

Indian Institute of Technology Roorkee https://orcid.org/0009-0008-7009-319X

Mayank Goswami

mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee

Research Article

Keywords: ATLAS, Higgs boson, reproducibility, exploratory data analysis, logistic regression, neural networks, approximate median significance, event weights.

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11633362/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.

Neural Network-Based Classification of Higgs Boson Signal and Background Events in the H$\to\tau\tau$ Channel Using ATLAS Open Data

Dhruv,[1] and Mayank Goswami,[1,2]

[1]Department of Physics, IIT Roorkee
[2]MF School of AI and DS, IIT Roorkee
mayank.goswami@ph.iitr.ac.in

Abstract*—The simulated ATLAS $H \to \tau\tau$ open dataset, released on the CERN Open Data Portal, is a convenient testbed for separating signal from background, but results obtained on its labelled file depend on preprocessing, event-weight, operating-point and uncertainty conventions that are easily left implicit. This paper does not propose a new classifier. Its contribution is a reproducible, leakage-controlled educational framework that combines gradient-verified NumPy implementations of logistic regression and a two-hidden-layer neural network, missingness-aware preprocessing fitted on training data only, and a frozen-protocol evaluation in which models are selected by validation weighted ROC-AUC, the selected-event fraction is fixed at 13.0% for every model, and the internal holdout is scored only after these choices are locked. Raw-subset approximate median significance (AMS, $b_{\mathrm{reg}} = 10$) is the primary yield metric; two $\times 5$ extrapolation conventions are reported separately as supplementary. On the 250,000 labelled simulated events a five-seed ensemble reaches a holdout weighted ROC-AUC of 0.9335 and a raw AMS of 1.56 (95% event-bootstrap interval [1.45, 1.67], 2,000 resamples); its individual members give AUC 0.9325 ± 0.0004 and raw AMS 1.51 ± 0.02 across seeds. The ensemble is clearly better than a tuned logistic-regression baseline (paired holdout AMS difference 0.85, interval [0.75, 0.96]), but its margin over tuned histogram gradient boosting (0.040, interval $[-0.032, 0.124]$) is not distinguishable from evaluation-sample noise and is not claimed as an improvement. Ablations are single-seed and directional. The data are simulated and no detector systematic uncertainties are modelled.

***Index Terms*—ATLAS, Higgs boson, $H \to \tau\tau$, reproducibility, exploratory data analysis, logistic regression, neural networks, approximate median significance, event weights.**

I. Introduction

The Higgs boson is the excitation of the Higgs field in the Standard Model. Its observation by the ATLAS and CMS experiments in 2012 [1], [2] established a major missing element of the model and opened a programme of precision measurements. The $H \to \tau^+\tau^-$ channel is important because it probes the Higgs coupling to charged leptons; at a Higgs mass near 125 GeV its Standard Model branching fraction is approximately 6.3% [3], [4].

Selecting such events is difficult. Many Standard Model processes produce final states that resemble a Higgs decay, and tau leptons decay with neutrinos that escape detection and appear only as missing transverse momentum. A useful classifier must therefore combine several correlated kinematic quantities. Machine learning has been applied to this kind of signal–background problem for some time [5], [6].

This study uses the simulated ATLAS $H \to \tau\tau$ dataset released on the CERN Open Data Portal [6], [7]. Anyone who evaluates a model on its labelled file must make several choices: where preprocessing statistics are fitted, how the -999 sentinel is encoded, how event weights are rescaled for a subset, how the selected-event operating point is chosen without touching the final evaluation sample, and how uncertainty is quantified. These choices change the reported numbers (Section IX) yet are easily left unstated.

A. Scope and contribution

Claim in one sentence. We present a reproducible, leakage-controlled educational framework for $H \to \tau\tau$ classification on the ATLAS open labelled data, and we report seed-wise, paired-bootstrap and directional ablation evidence on how its components change validation and holdout AUC and AMS (the ablations are single-seed and budget-confounded, Section IX); we do not claim a new classifier.

The specific contributions are:

- C1 **A leakage-controlled protocol.** Stratified 60/20/20 splitting; preprocessing statistics and class-normalisation factors fitted on the training partition only; model selection by validation weighted ROC-AUC; one selected-event fraction ($f^* = 13.0\%$) applied unchanged to every model, seed and partition; holdout scored only after these choices are frozen (Fig. 4).
- C2 **Missingness-aware preprocessing with directional evidence.** Explicit indicators for the structural -999 sentinel, invalid-value indicators and $\log(1 + x)$ transforms for 14 non-negative variables, $\sin / \cos$ encoding of azimuthal angles and a four-category jet encoding (63 inputs), with single-seed ablations of each element (Table VII); these are not clean one-factor tests.
- C3 **Gradient-verified from-scratch implementations.** Logistic regression and a two-hidden-layer network

Corresponding author: mayank.goswami@ph.iitr.ac.in

in NumPy. A finite-difference check (maximum relative error 6.55×10^{-10}) covers the unweighted, unregularised gradient of a small one-hidden-layer test network; the weighted, L2 and two-layer paths are not separately tested.

C4 **Transparent evaluation conventions, uncertainty and a claims ledger.** Raw-subset AMS is primary; two $\times 5$ yield conventions are reported separately and their exact relation is derived; seed-wise intervals, a 2,000-resample event bootstrap and paired bootstrap differences are reported; Table VIII states which conclusions the evidence does and does not support.

B. Research questions

RQ1: Which preprocessing and objective choices appear to change validation AUC and AMS, and how far can the recorded runs support that? **RQ2:** How do the weighting and AMS conventions affect which model appears best? **RQ3:** Which headline comparisons survive an honest accounting of uncertainty?

The rest of the paper is organised as follows. Section II gives background and states the gap. Sections III and IV describe the dataset and exploratory analysis. Sections V–VIII give the methods. Sections IX and X present results and discussion, and Section XI lists limitations and reproducibility information.

II. Background and Gap

Machine learning in high-energy physics. Machine learning is widely applied to signal–background separation in collider physics; for example, Baldi et al. showed that deep networks can recover discriminating power from low-level features on other exotic-particle datasets [5]. The dataset used here, together with its documentation and the AMS figure of merit, is described in [6], [7].

Scope of comparisons. All model comparisons in this paper are made on the same split, with the same four-category engineered feature matrix, the same training-derived preprocessing, the same class-normalised training weights, the same original evaluation weights and the same fixed top-fraction rule (Table VI). Logistic regression and histogram gradient boosting [8] are tuned over small, explicitly reported four-point grids on validation weighted AUC; the NumPy networks represent the neural-network family. Absolute AMS values depend on the sample definition and weight normalisation and are not compared here with values obtained under other conventions.

Gap addressed. What a student or practitioner reusing the labelled file needs is a documented protocol that (i) keeps preprocessing and operating-point choice strictly out of the evaluation sample, (ii) states how subset event weights are rescaled and what that does to AMS, and (iii) reports which comparisons are statistically distinguishable and which uncertainty each interval does and does not contain. This paper supplies and exercises such a protocol; it does not claim to fill a gap in classifier accuracy.

TABLE I
Labelled-sample composition.

Class	Events	Fraction
Background (b)	164,333	65.73%
Signal (s)	85,667	34.27%
Total	250,000	100.00%

III. Physics Context and Dataset

A. Detector quantities and target process

ATLAS is a general-purpose detector at the Large Hadron Collider. Tracking detectors measure charged-particle trajectories, calorimeters sample energy deposits, and the muon system identifies and measures muons; these measurements are reconstructed into physics objects and derived event-level quantities [9]. In this dataset, names beginning with `PRI_` denote primitive reconstructed quantities and `DER_` derived kinematic quantities, including transverse momenta (p_T), pseudorapidity (η), azimuthal angle (ϕ), missing transverse energy, invariant-mass estimates and topology variables.

The label is `s` (signal) or `b` (background), encoded as 1 and 0. Signal is simulated $H \to \tau\tau$; background combines other simulated processes that can mimic the selected final state. The dataset derives from official ATLAS full-detector simulation, in which each event's origin is known [7]. It is therefore a supervised benchmark on simulation, not an analysis of recorded collisions and not a discovery claim.

B. Dataset structure

The labelled file contains 250,000 events and 33 columns: an identifier, 30 input features, a per-event `Weight` and the label. The unlabelled test file used here has 550,000 events with identifier and features only (no labels or weights), so classification performance on it cannot be computed locally. `Weight` is not a detector feature and is never an input to any model here. Table I gives the class composition.

C. Event weights and the target metric

Raw class proportions do not equal the expected physical composition. The summed signal weight in the labelled sample is only about 692, against roughly 411,000 for background, so accuracy is not an adequate final metric. (This figure is quoted in the EDA notebook; the $\times 5$-rescaled holdout weights in Eq. (21) sum to about 694.6 for signal and 410,386 for background, consistent with it.) The figure of merit used here is the approximate median significance (AMS), defined in Section VIII, which combines the weighted signal yield s and weighted background yield b of a selected event set [6]. AMS suits benchmark ranking but is not a discovery significance: published ATLAS analyses use likelihood models with signal and control regions, nuisance parameters and profile-likelihood test statistics [10], [11].

IV. Exploratory Data Analysis

A. Sentinel values and structural missingness

Unavailable quantities are stored as -999, so the file has no conventional missing entries at load time. Treating the sentinel as missing, the EDA finds about 70.98% missingness across seven dijet/subleading-jet variables, 39.97% across three leading-jet variables and 15.25% in `DER_mass_MMC` (Fig. 1). These are not random sensor failures: a dijet mass is undefined when an event has fewer than two reconstructed jets. A missingness indicator can therefore carry information about event topology.

B. Jet multiplicity and signal purity

`PRI_jet_num` takes values 0, 1, 2 and 3 (3 meaning three or more jets). Signal purity is about 25.5% for zero-jet, 35.7% for one-jet, 51.1% for two-jet and 30.4% for three-or-more-jet events, so jet topology is informative and also explains which variables are undefined. Jet multiplicity is therefore encoded categorically rather than as a numeric scale; because the two-jet and three-or-more-jet purities differ markedly, the four categories are kept separate.

C. Skewness, tails and correlations

Several energy and mass features are positively skewed. For `DER_mass_MMC` the skewness is 5.41 for signal, 3.23 for background and 3.81 combined, so a symmetric Gaussian is a poor description. A few features, including `DER_pt_tot`, show extreme skewness and kurtosis driven by individual outliers; this should not be read as a general property of all events. Feature-pair correlations near $+1$ that involve variables sharing the same sentinel pattern may reflect shared structural missingness rather than a physical linear relation. Replacing the sentinel without keeping its indicator would hide this distinction, while leaving it numeric would create a spurious magnitude; the pipeline therefore creates explicit indicators before imputation.

Marginal histograms of training and official test features look broadly similar. This is a useful sanity check but does not prove equal multivariate distributions or exclude covariate shift. Fig. 2 shows that signal and background overlap in the `DER_mass_MMC`–`DER_mass_jet_jet` projection but that signal concentrates in a narrower reconstructed-mass region; this is a projection of simulated events, not evidence of a resonance in any single event.

D. Feature distributions and train–test comparison

Fig. 3 overlays training and test histograms for the 29 continuous features (the categorical `PRI_jet_num` is omitted; sentinel values are excluded and 81 bins are used). The test file holds about 2.2 times more events than the training file, so the grey histograms are taller; apart from this scale difference the training and test shapes agree for every feature, as expected for samples from the same simulation. This is a marginal, visual check only. It cannot detect differences in joint distributions, and tail-sensitive statistics can still differ: the skewness of `DER_pt_tot` is 10.58 in the training file and 2.85 in the test file, which likely reflects a few extreme outliers. A more rigorous test, such as a classifier trained to separate training from test rows, was not performed. The features fall into five families.

Energy, mass and momentum features (`DER_mass_MMC`, `DER_mass_vis`, `DER_mass_transverse_met_lep`, `DER_pt_h`, `DER_pt_tot`, `DER_sum_pt`, `DER_mass_jet_jet`, `DER_pt_ratio_lep_tau`, and the `PRI_` transverse-momentum and `PRI_met` variables) show a sharp peak at low values and a long right tail, a falling spectrum in which high-energy collisions are much rarer than low-energy ones. A few events reach values of order 10^3 GeV or more (up to about 5,000 GeV for `DER_mass_jet_jet`), which compresses the visible range; these variables are strongly right-skewed and heavy-tailed (Section IV-C). `DER_mass_MMC` peaks at a value of order 100 GeV.

Pseudorapidity features (`PRI_tau_eta`, `PRI_lep_eta`, `PRI_jet_leading_eta`, `PRI_jet_subleading_eta`) are symmetric about zero, with sharp cut-offs near ± 2.5 for the tau and lepton (consistent with tracker acceptance) and near ± 4.5 for jets (consistent with calorimeter coverage); the symmetry reflects the two identical proton beams.

Azimuthal-angle features (the five `PRI_*_phi` variables) are approximately uniform on $[-\pi, \pi]$, since no transverse direction around the beam is special. Azimuth alone therefore carries little information; only angle differences, as in ΔR or the centrality variables, are informative.

Bounded shape variables. `DER_met_phi_centrality` lies in $[-\sqrt{2}, \sqrt{2}]$ with pronounced peaks near both bounds, the upper one corresponding to missing transverse momentum pointing between the tau and lepton directions. `DER_lep_eta_centrality` lies in $[0, 1]$ with peaks at 0 (lepton far outside the two leading jets) and 1 (lepton midway between them). `DER_deltar_tau_lep` rises to a peak near π (tau and lepton back to back) and falls off, with a longer tail towards small values.

Jet-pair variables. `DER_deltaeta_jet_jet` is an absolute difference, so it starts at zero and falls towards about 8.5. `DER_prodeta_jet_jet` is a product of two pseudorapidities, symmetric about zero and sharply peaked there, with wide tails (roughly -20 to $+17$ in these data). Both are defined only for events with at least two jets, about 29% of the sample.

V. Framework and First-Principles Logistic Regression

A. Framework overview

Fig. 4 summarises the protocol. Its defining rules are: the labelled data are split once, stratified by label; every preprocessing statistic and class-normalisation factor is computed from the training partition; all model selection (network configuration, logistic regularisation strength, boosting settings) uses validation weighted ROC-AUC; one selected-event fraction, $f^* = 13.0\%$, is applied to every model, seed

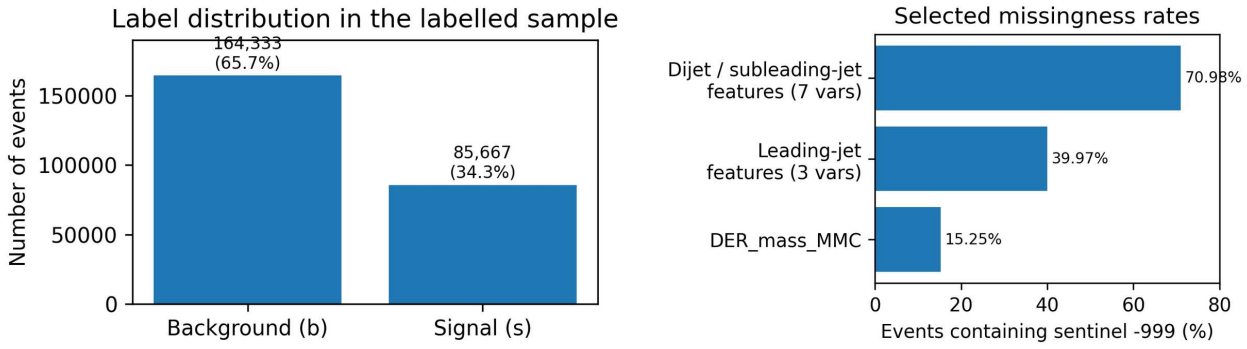


Fig. 1. EDA summary. (a) Signal/background counts in the 250,000-event labelled sample. (b) Sentinel (-999) rates in selected feature families. Label imbalance is moderate; jet-related missingness is much larger and largely structural.

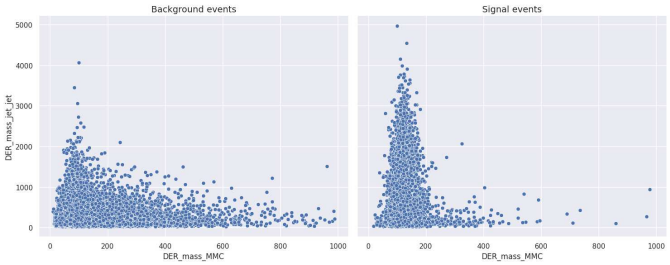


Fig. 2. `DER_mass_MMC` versus `DER_mass_jet_jet` for background (left) and signal (right) simulated events. The distributions overlap; signal is more concentrated in reconstructed mass.

and partition and is never searched on the holdout; the holdout is scored after these choices are frozen; and the official test file is used only to produce a prediction file.

B. Model and probability interpretation

Let $X \in \mathbb{R}^{m\times n}$ hold m events and n features, $\mathbf{w} \in \mathbb{R}^n$ the coefficients and b the intercept. The linear score and estimated signal probability are

$$z_i = \mathbf{x}_i^T\mathbf{w} + b, \tag{1}$$

$$p_i = \sigma(z_i) = \frac{1}{1+e^{-z_i}}. \tag{2}$$

The decision rule at threshold t is $\hat{y}_i = 1$ if $p_i \geq t$, else 0. At $t = 0.5$ the boundary is $\mathbf{x}^T\mathbf{w} + b = 0$.

C. Binary cross-entropy and its gradient

For $y_i \in \{0, 1\}$ the negative log-likelihood is $\ell_i = -y_i \ln p_i - (1 - y_i)\ln(1 - p_i)$ and the mean objective is

$$J(\mathbf{w}, b) = -\frac{1}{m}\sum_{i=1}^{m}\big[y_i \ln p_i + (1 - y_i)\ln(1 - p_i)\big]. \tag{3}$$

Using $\partial p_i/\partial z_i = p_i(1 - p_i)$ gives $\partial\ell_i/\partial z_i = p_i - y_i$, hence

$$\nabla_{\mathbf{w}} J = \frac{1}{m}X^T(\mathbf{p} - \mathbf{y}), \qquad \frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}(p_i - y_i). \tag{4}$$

These were coded both as explicit loops and as vectorised NumPy, and the two agree under an all-close check. A two-observation toy test gives $\nabla_{\mathbf{w}} J \approx (-4.99991649, 4.99991649)$ and $\partial J/\partial b \approx 0.49999165$; this is a derivative test, not a Higgs-trained result.

D. Batch gradient descent

With learning rate α each full-batch update is

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha\nabla_{\mathbf{w}} J, \qquad b^{(k+1)} = b^{(k)} - \alpha\,\partial J/\partial b. \tag{5}$$

The Higgs logistic notebook starts from zero weights and bias, uses $\alpha = 0.1$ and runs 2,000 updates. The training objective falls from 0.687174 at the first reported update to 0.592539 at update 2,000, an absolute reduction of 0.094635 (Fig. 5); it is still decreasing at update 2,000, so convergence is not demonstrated.

E. Historical exploratory run and its methodological flaw

The original logistic-regression run is retained as a historical educational baseline. It removed ten high-missingness jet features (leaving 20), imputed the remaining missing mass feature, used a non-stratified 80/20 split, computed the median imputation value separately in each subset, and fitted min–max scaling independently on the training and holdout subsets. Independent scaling of the holdout is a methodological flaw: the fitted coefficients act on a different feature scale at test time, and the holdout influences preprocessing. The revised pipeline (Section VI) fits all statistics on the training partition alone. The size of the effect of this flaw on the original numbers was not isolated, because the revised pipeline also changed the split and feature set; the original numbers are therefore reported as a record, not as a controlled benchmark. Only the train-fitted logistic baseline of Section IX-C is a fair same-split reference.

At threshold 0.5 the original run gives 67.79% training and 68.34% holdout accuracy. The holdout confusion matrix (rows true [b, s], columns predicted [b, s]) is

$$C_{\mathrm{LR}} = \begin{pmatrix} 28722 & 4089 \\ 11739 & 5450 \end{pmatrix}. \tag{6}$$

It implies signal precision $5450/(5450 + 4089) = 57.1\%$, signal recall $5450/(5450 + 11739) = 31.7\%$, background

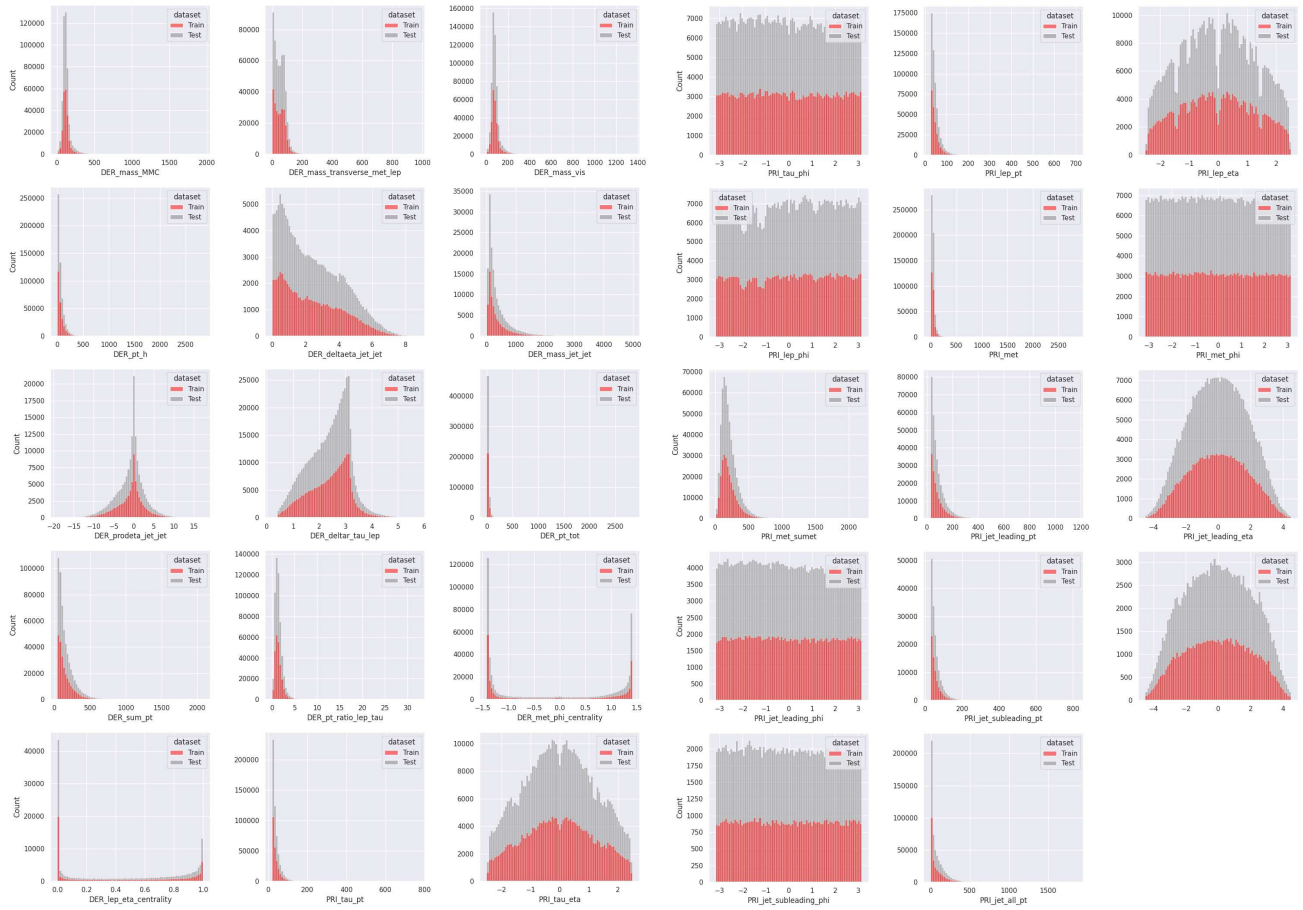


Fig. 3. Training (red) and test (grey) histograms of the 29 continuous features (81 bins, event counts, sentinel values excluded). The test file contains about 2.2 times more events, so grey bars are taller. Panels follow the column order of the data file, the left three columns first. The categorical `PRI_jet_num` is not shown.

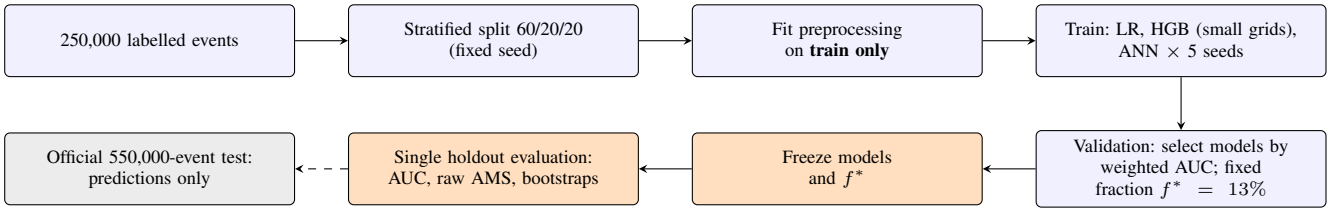


Fig. 4. Leakage-controlled evaluation protocol. Orange boxes mark the frozen stage: nothing fitted or chosen after this point may use holdout information.

specificity $28722/(28722 + 4089) = 87.5\%$ and balanced accuracy about 59.6% (Table II). The low recall shows why accuracy alone is inadequate: the classifier handles most background but misses most signal. Indeed, always predicting background would be correct for $32{,}811/50{,}000 = 65.62\%$ of this holdout, so the 68.34% accuracy is only 2.72 percentage points above a trivial rule.

VI. Preprocessing for the Revised Pipeline

The labelled rows are split by a stratified 60/20/20 rule into 150,000 training, 50,000 validation and 50,000 internal-holdout events; the signal fraction is 0.3427 in each. `EventId`, `Weight` and `Label` are excluded from the predictors. All statistics below are fitted on the training partition and stored.

Missingness flags. Eleven variables get a flag $m_j = \mathbf{1}[x_j = -999]$ computed before the sentinel is converted to missing and imputed: `DER_mass_MMC`, four dijet/centrality variables and six leading/subleading-jet kinematic variables.

Log transform and invalid-value flags. Fourteen predefined, physically non-negative, right-skewed kinematic inputs (masses, transverse momenta, scalar sums and missing-energy variables) are transformed by $x'_j = \ln(1+x_j)$ after the sentinel is removed. A negative observed value in one of these variables would be physically exceptional; it is not silently clipped but set to missing and recorded by a separate invalid-value flag.

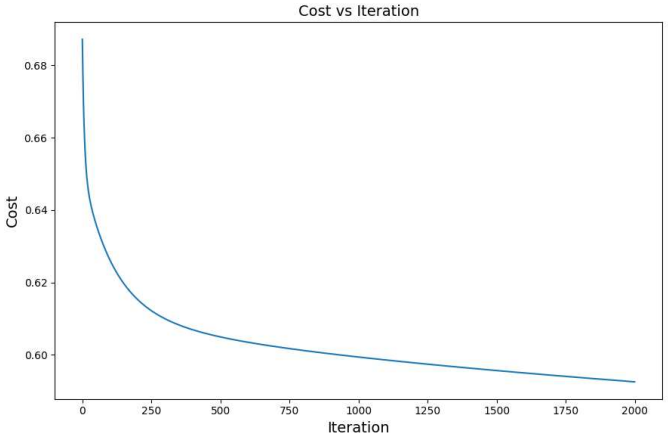


Fig. 5. Training binary cross-entropy versus iteration for the original first-principles logistic-regression run. The decrease confirms optimisation of the training objective; it does not by itself establish generalisation.

TABLE II
ORIGINAL FIRST-PRINCIPLES LOGISTIC-REGRESSION RESULTS (HISTORICAL RUN).

Quantity	Recorded result
Gradient-descent updates	2,000
Learning rate	0.1
Training BCE, first reported update	0.687174
Training BCE, update 2,000	0.592539
Training accuracy, threshold 0.5	67.79%
Holdout accuracy, threshold 0.5	68.34%
Signal precision	57.1%
Signal recall	31.7%
Background specificity	87.5%
Balanced accuracy	59.6%

One such flag column exists for each of the 14 variables. Because these variables are non-negative by construction in this dataset, these columns are expected to be constant zero in the training data; this was not printed in the recorded run, so they are counted below as part of the 63 inputs but are not credited with any contribution.

Angles. Five azimuthal variables are represented by $\sin\phi$ and $\cos\phi$, avoiding the discontinuity between $-\pi$ and $+\pi$.

Jet multiplicity. Jet number is one-hot encoded in four mutually exclusive categories: 0, 1, 2 and $\geq$ 3 jets (the source value 3 denotes three or more). This preserves the purity difference between two-jet (51.1%) and three-or-more-jet (30.4%) events.

Feature count. The representation has $30-1-5+10+4+11+14=63$ columns (30 raw inputs; minus `PRI_jet_num`; minus five azimuths plus ten sine/cosine columns; four jet categories; 11 missingness flags; 14 invalid-value flags). All reported ANN, logistic-regression and boosting results use this same 63-column matrix.

Standardisation. Medians, means and standard deviations come from the training partition only:

$$z_{ij} = \frac{x'_{ij} - \mu_{j,\text{train}}}{\sigma_{j,\text{train}}}, \tag{7}$$

with zero-variance safeguards, and the same values and feature order are applied unchanged to validation, holdout and official test data.

VII. FROM-SCRATCH NEURAL NETWORK

A. Architecture

The network has hidden layers of 64 and 32 ReLU units and one sigmoid output. With $\mathbf{a}^{(0)} = \mathbf{x}$ (63 inputs),

$$\mathbf{z}^{(\ell)} = W^{(\ell)}\mathbf{a}^{(\ell-1)} + \mathbf{b}^{(\ell)}, \quad \mathbf{a}^{(\ell)} = \max(0, \mathbf{z}^{(\ell)}),\ \ell = 1, 2, \tag{8}$$

$$\hat{p} = \sigma\big(W^{(3)}\mathbf{a}^{(2)} + b^{(3)}\big). \tag{9}$$

Hidden weights use He-style initialisation [12]; the output layer uses a smaller fan-in-based scale. A numerically stable sigmoid and probability clipping prevent overflow and $\log 0$. Each member has 6,209 parameters.

B. Weighted objective and regularisation

Training uses a class-normalised event-weighted cross-entropy. Let w_i be the supplied weight and $c_i \in \{0, 1\}$ the class. Training weights are $\tilde{w}_i = w_i N/(2W_{c_i})$, where N is the number of training events and W_c the summed raw training weight of class c, so that signal and background contribute equally and the mean weight is 1; the factors are fitted on the training split only and applied unchanged to the validation loss used for early stopping. The objective is

$$J(\theta) = -\frac{\sum_i \tilde{w}_i \left[y_i \ln \hat{p}_i + (1-y_i)\ln(1-\hat{p}_i)\right]}{\sum_i \tilde{w}_i} + \frac{\lambda}{2}\sum_{\ell} \|W^{(\ell)}\|_F^2. \tag{10}$$

Gaussian input noise with standard deviation 0.02 is added during training. Evaluation always uses the *original* event weights; class normalisation is a training choice and does not alter the evaluation yields or the weighted AUC.

C. Backpropagation and gradient verification

For a ReLU unit the derivative is $\mathbf{1}[z > 0]$; for a sigmoid output with cross-entropy the derivative with respect to the logit reduces to $\hat{p} - y$. Gradients take the usual form

$$\frac{\partial J}{\partial W^{(\ell)}} = \Delta^{(\ell)}\big(A^{(\ell-1)}\big)^T + \lambda W^{(\ell)}, \qquad \frac{\partial J}{\partial \mathbf{b}^{(\ell)}} = \sum_i \Delta_i^{(\ell)}, \tag{11}$$

with mini-batch averaging in the implementation. A finite-difference check gives a maximum relative gradient error of 6.552×10^{-10} and passes. The test network has four inputs, one hidden layer of three units, unweighted loss and $\lambda = 0$, so it verifies the core chain-rule implementation but does not separately test the event-weight scaling, the L2 term or the two-hidden-layer path. It is a correctness test, not a performance metric.

TABLE III
HYPERPARAMETER SEARCH (SINGLE-SEED RUNS ON A 40,000-ROW TRAINING SUBSET, ≤35 EPOCHS). SELECTION BY VALIDATION WEIGHTED AUC. RAW AMS IS AT THE FIXED 13% FRACTION.

Cfg.	Hidden units	LR	L2	Val. AUC	Raw AMS
1	(32,16)	10^{-3}	10^{-4}	0.930413	1.5170
2	(64,32)	10^{-3}	10^{-4}	0.930440	1.4680
3	(64,32)	5×10^{-4}	10^{-4}	0.927788	1.3736
4	(64,32)	10^{-3}	10^{-3}	**0.931337**	1.4907

D. Adam optimisation, early stopping and ensembling

Adam [13] keeps first and second moment estimates

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \tag{12}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \tag{13}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}, \tag{14}$$

$$\theta_{t+1} = \theta_t - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}. \tag{15}$$

The settings are $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, initial rate $\eta_0 = 10^{-3}$, batch size 512, at most 80 epochs per seed, and schedule $\eta_e = \eta_0/[1 + 0.01(e-1)]$. Each seed uses early stopping on validation loss with patience 10 and restores its best checkpoint. Five models (seeds 243–247) are averaged:

$$\hat{p}_{\text{ens}}(\mathbf{x}) = \frac{1}{5} \sum_{k=1}^{5} \hat{p}_k(\mathbf{x}). \tag{16}$$

Ensembling reduces sensitivity to a single initialisation but does not remove sampling or model-selection uncertainty. Each member's predictions, and each member's validation and holdout metrics, are retained.

E. Validation hyperparameter search

Four configurations were compared on the same validation split (Table III). To keep the search quick, each candidate was trained on a fixed stratified 40,000-row subset of the training split for at most 35 epochs (patience 7), whereas the final ensemble members use all 150,000 training rows and up to 80 epochs. **Validation weighted ROC-AUC is the model-selection objective**, because the retained model is intended as a global event-ranking classifier; raw AMS at the fixed fraction is reported as a secondary diagnostic only. Configuration 4, hidden units $(64, 32)$ with learning rate 10^{-3} and L2 coefficient 10^{-3}, has the highest AUC. It does not have the highest raw AMS (configuration 1 does), and the AUC differences among configurations 1, 2 and 4 are at most 0.0009 in single-seed runs on reduced data, several of which stopped within a few epochs of the epoch cap. The choice among these near-ties is therefore a convention, not an established optimum, and selection uncertainty is not quantified.

VIII. EVALUATION PROTOCOL

A. Weighted yields and AMS

For a selected set of events, the weighted signal and background yields are

$$s = \sum_i w_i \, \mathbf{1}[y_i = 1] \, \mathbf{1}[\hat{y}_i = 1], \tag{17}$$

$$b = \sum_i w_i \, \mathbf{1}[y_i = 0] \, \mathbf{1}[\hat{y}_i = 1], \tag{18}$$

and the AMS is

$$\text{AMS}(s, b; b_{\text{reg}}) = \sqrt{2\Big[(s + b + b_{\text{reg}}) \ln\Big(1 + \frac{s}{b + b_{\text{reg}}}\Big) - s\Big]}, \tag{19}$$

with $b_{\text{reg}} = 10$ following the definition in the dataset documentation [6]. It is a benchmark objective, not a significance that includes systematic uncertainties. Direct-formula unit tests and a homogeneity test (below) are included in the notebook.

B. Weight conventions and what they do to AMS

Three quantities are distinguished.

1) **Raw-subset AMS (primary).** Computed from the actual 50,000-event validation or holdout subset, with the original event weights and $b_{\text{reg}} = 10$. It is reproducible directly from scores, labels, weights and the fixed fraction.
2) **Yield-extrapolated AMS, fixed regulator (supplementary heuristic).** Both yields are multiplied by $r = N_{\text{labelled}}/N_{\text{subset}} = 250000/50000 = 5$ and the formula is evaluated with $b_{\text{reg}} = 10$, i.e. $\text{AMS}(rs, rb; 10)$. This is not $r \times \text{AMS}(s, b; 10)$; it changes the relative weight of the regulator and is not an official score.
3) **Scale-consistent sensitivity convention.** The regulator is scaled with the yields, $\text{AMS}(rs, rb; r\, b_{\text{reg}})$. Since the bracket in the AMS formula is homogeneous of degree one in (s, b, b_{reg}),

$$\text{AMS}(rs, rb; r\, b_{\text{reg}}) = \sqrt{r}\, \text{AMS}(s, b; b_{\text{reg}}), \tag{20}$$

so this convention adds no information beyond the raw value (holdout: $1.555207 \times \sqrt{5} = 3.477549$, matching the computed value exactly). It is a mathematical sensitivity check, not the competition metric.

None of these is the dataset's own class-wise weight normalisation [6], [7], and none may be compared with significance values from other sample definitions unless the sample, weight normalisation, selected-event rule and regulator are all aligned. In this paper they are used only for relative comparisons between models evaluated identically.

C. Operating point

Events are ranked by model score and a fixed top fraction is selected rather than applying a raw probability cutoff. The fraction is fixed at $f^* = 13.0\%$ for every model, seed and partition; the corresponding score cutoff differs between partitions (0.9505 on validation, 0.9501 on the holdout) because the score distributions differ. The value 13.0% is carried over

from an earlier iteration of this pipeline, in which it was the validation-AMS maximiser for an earlier ensemble; it was not chosen on the holdout and is not re-optimised here. The full validation AMS-versus-fraction curves are shown only as diagnostics (Fig. 7). In the present runs the individual seeds peak at 14.5–16.0% on validation, so 13% sits slightly below the plateau and costs between about 0.02 and 0.08 raw AMS per seed; it is a fixed convention, not a claimed optimum.

D. Uncertainty

Three different sources of variability are reported separately and must not be conflated. (i) *Seed-wise variability:* the five ensemble members are individually evaluated; means, sample standard deviations and 95% Student-t intervals ($t_{0.975,4} = 2.776$) are reported. (ii) *Evaluation-sample variability:* percentile intervals from 2,000 nonparametric event-level bootstrap resamples [14], with f^* held fixed, and paired bootstrap differences in which two models are scored on identical resampled events. (iii) *Omitted sources:* retraining variability of the whole pipeline, hyperparameter-selection uncertainty and detector systematics are not captured by either interval. Single-seed models (logistic regression, boosting, the shallow network and every ablation) have no seed-wise interval.

E. Calibration and importance diagnostics

Calibration is summarised by the weighted Brier score and a ten-bin weighted expected calibration error (ECE) [15]. Permutation importance measures the drop in validation weighted AUC when one engineered column is permuted.

IX. Results

A. Logistic regression

The first-principles regression demonstrates the full gradient-based optimisation path (Section V); its training objective decreases monotonically from 0.687174 to 0.592539. The historical holdout results (68.34% accuracy, 31.7% recall, 59.6% balanced accuracy; Table II) are unweighted statistics from the 20-feature pipeline. The 11,739 missed signal events outnumber the 5,450 correct detections by more than two to one, illustrating the limits of a single linear surface at this threshold. The controlled logistic baseline of Section IX-C uses the same split, features and weights as the networks.

B. ANN performance

The five-seed ensemble reaches validation weighted ROC-AUC 0.937088 and holdout weighted ROC-AUC 0.933462 (Table IV, Fig. 6). At the fixed fraction of 13.0% the validation raw AMS is 1.604310 and the holdout raw AMS is 1.555207; under the two supplementary $\times 5$ conventions the holdout values are 3.500703 ($b_{\mathrm{reg}} = 10$) and 3.477549 ($b_{\mathrm{reg}} = 50$), which differ from $\sqrt{5}\times$raw only through the regulator (Eq. (20)). Holdout weighted yields are $s = 38.313530$, $b = 584.276898$ before scaling and $s = 191.567651$, $b = 2921.384492$ after; validation yields are $s = 38.324287$, $b = 548.019159$ before and $s = 191.621435$, $b = 2740.095797$ after.

Seed-wise results. Table V gives each member's metrics. Across the five seeds the validation weighted AUC is 0.936131 ± 0.000175 (mean $\pm$ sample SD; 95% t-interval 0.935915–0.936348) and the validation raw AMS is 1.5796 ± 0.0390 (1.531–1.628); on the holdout the corresponding values are 0.932513 ± 0.000383 (0.932038–0.932989) and 1.5109 ± 0.0229 (1.482–1.539). The ensemble's AUC exceeds every member's on both partitions, by about 0.001; its AMS exceeds every member on the holdout but not on validation (seed 244 reaches 1.6407 there), so the AMS benefit of ensembling is not established. Best epochs are 40, 45, 69, 31 and 37 (mean 44.4, SD 14.7), showing considerable variability in stopping time.

Selected sample. On the rescaled holdout the weighted confusion matrix (rows true b/s, columns predicted b/s) is

$$C_{\mathrm{ANN}}^{\mathrm{holdout}} = \begin{pmatrix} 407465.0 & 2921.4 \\ 503.1 & 191.6 \end{pmatrix}, \tag{21}$$

entries being weighted, extrapolated yields rather than event counts. The selection retains 27.6% of the signal weight while keeping 0.71% of the background weight (validation: 27.8% and 0.67%; Fig. 8). Because a large absolute background weight remains in the selected set (2,740 versus 192 for signal on validation), the weighted signal fraction $s/(s+b)$ of the selection is only about 6.5% on validation and 6.2% on the holdout. This is why AMS, which depends on s and b together, rather than accuracy is the figure of merit used here.

Calibration and score distributions. Calibration is weak: the validation weighted Brier score is 0.112636 and the ten-bin ECE is 0.204453. Class normalisation changes the effective class prior, so the raw outputs are ranking scores rather than calibrated posteriors. If calibrated probabilities are needed, Platt scaling [16] or isotonic regression [17] should be fitted on independent calibration data or out-of-fold predictions, not on the data used to evaluate them. Fig. 9 shows how the ensemble score separates the two labels on validation (50 bins, each class normalised to unit area): background concentrates near zero and signal near one, the dashed line marks the score cutoff (0.9505) of the selected 13%, and the two classes overlap over a wide range of intermediate scores. These distributions describe the sample; because each class is normalised separately they do not show the physical class proportions.

Evaluation-sample uncertainty. Fig. 10 shows the bootstrap distributions of the raw AMS. The holdout interval ([1.454, 1.668]) lies lower than the validation interval ([1.503, 1.750]) and the two overlap. The lower holdout value is consistent with the optimism expected from stopping epoch and configuration being chosen on validation data and with sampling variation; because the fraction is fixed rather than maximised, optimism from operating-point search is largely absent. These intervals quantify evaluation-sample variability only. As a diagnostic of how sharply the operating point is determined, a 500-resample bootstrap of the validation-AMS peak location (coarse grid up to 20%) gives a mean of 15.5%

TABLE IV

ENSEMBLE EVALUATION AT THE FIXED 13.0% FRACTION. RAW AMS IS PRIMARY ($b_{\text{reg}} = 10$); THE $\times 5$ COLUMNS ARE SUPPLEMENTARY CONVENTIONS. INTERVALS ARE 95% PERCENTILE INTERVALS FROM 2,000 EVENT-LEVEL BOOTSTRAP RESAMPLES (FRACTION HELD FIXED).

Split	Weighted AUC	Fraction	Raw AMS	95% CI, raw AMS	$\times 5$ AMS ($b_{\text{reg}} = 10$)	$\times 5$ AMS ($b_{\text{reg}} = 50$)	Score cutoff
Validation	0.937088	13.0%	1.604310	[1.503468, 1.749893]	3.612766	3.587347	0.950478
Internal holdout	0.933462	13.0%	1.555207	[1.454406, 1.667898]	3.500703	3.477549	0.950051

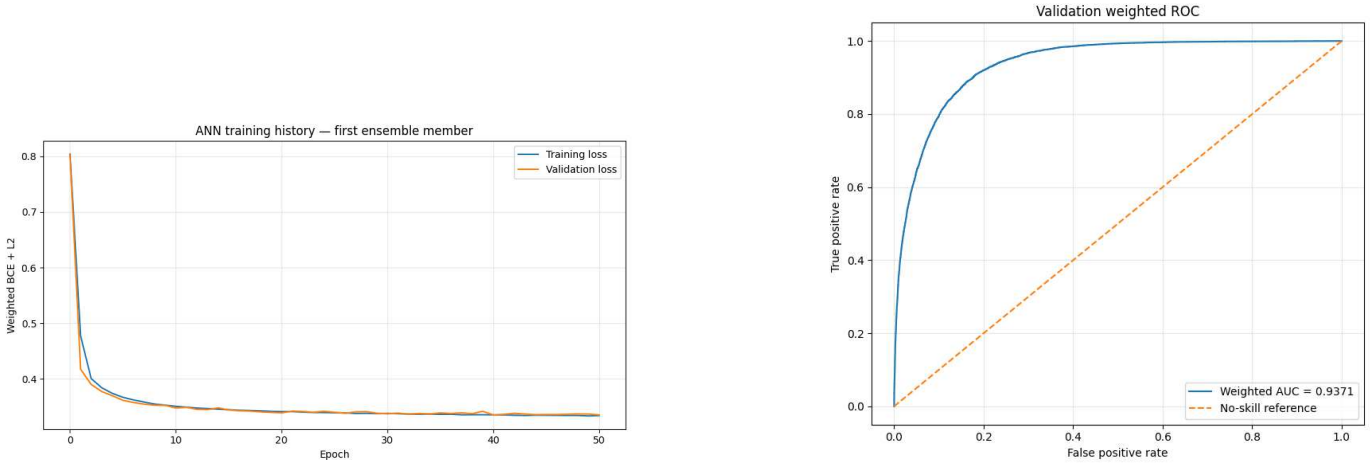


Fig. 6. NumPy ANN diagnostics. (a) Training and validation weighted objective for the first ensemble member: rapid early improvement then a flat trajectory. (b) Validation ROC curve of the ensemble with original event weights; weighted AUC 0.9371.

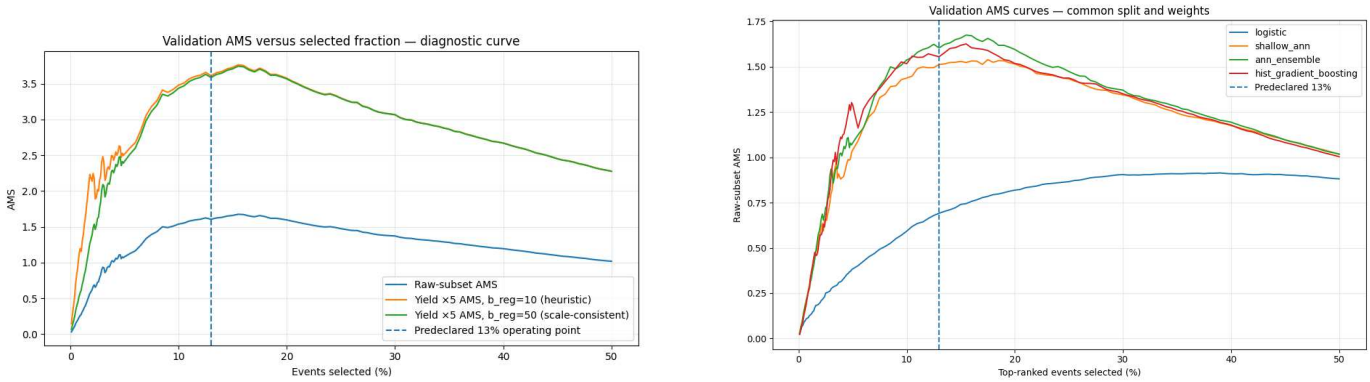


Fig. 7. Diagnostic validation AMS versus the fraction of top-ranked events selected. (a) Ensemble: raw-subset AMS and the two $\times 5$ conventions; the dashed line marks the fixed 13% fraction, which is not chosen from this curve. (b) Raw-subset AMS for the four principal models on the same split and weights; the logistic curve peaks only at much larger fractions. The curves are noisy at small fractions.

(SD 1.5%; 5th–95th percentile 12–18%): the AMS curve has a broad, noisy plateau, not a sharp optimum.

C. Controlled comparison on the same split

Table VI compares score functions on the same validation and holdout splits, the same 63-column feature matrix, the same training weights, the same evaluation weights and regulator, and the same fixed 13% fraction. Logistic regression (scikit-learn L-BFGS, up to 1,000 iterations) was tuned over $C \in \{0.01, 0.1, 1, 10\}$ (best $C = 10$; validation AUC varies only between 0.8828 and 0.8831). Histogram gradient boosting was tuned over four settings (best: 100 iterations, learning rate 0.08, 31 leaves, L2 1.0; validation AUC 0.9348, versus 0.9307–0.9347 for the others). The shallow ANN has one hidden layer of 16 units and was capped at 50 epochs. The ANN was tuned over four configurations and is a five-model ensemble, whereas boosting and logistic regression are single fits, so the table does not place the ANN family and the boosting family on an equal footing.

Versus logistic regression. The ensemble–logistic gap is large (validation AUC $+0.0540$; holdout raw AMS paired difference, ANN minus logistic: median 0.851, 95% percentile interval [0.753, 0.960] from 2,000 paired resamples). Part of the AMS gap reflects the common fixed fraction: the logistic curve peaks at much larger fractions (Fig. 7b), so at 13% it operates away from its own optimum. The AUC gap does not

TABLE V

SEED-WISE ANN RESULTS AT THE FIXED 13% FRACTION (RAW AMS, $b_{\mathrm{reg}} = 10$). SD IS THE SAMPLE STANDARD DEVIATION; INTERVALS ARE 95% STUDENT-t ($df = 4$).

Seed	Best ep.	Val. AUC	Val. AMS	Hold. AUC	Hold. AMS
243	40	0.936068	1.5414	0.932704	1.4798
244	45	0.936220	1.6407	0.932695	1.5195
245	69	0.936320	1.5503	0.932917	1.4969
246	31	0.935864	1.5846	0.932287	1.5392
247	37	0.936184	1.5806	0.931963	1.5190
Mean	44.4	0.936131	1.5796	0.932513	1.5109
SD	14.7	0.000175	0.0390	0.000383	0.0229
95% interval	–	0.9359–0.9363	1.531–1.628	0.9320–0.9330	1.482–1.539
Ensemble	–	0.937088	1.6043	0.933462	1.5552

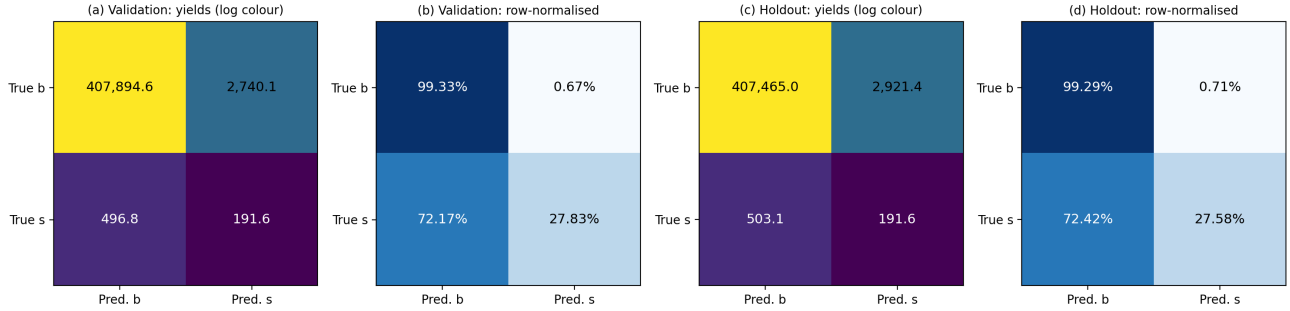


Fig. 8. Weighted confusion matrices at the fixed 13.0% fraction, with $\times 5$ rescaled weights; entries are weighted yields, not event counts. (a), (b) Validation; (c), (d) internal holdout. Panels (a) and (c) use a logarithmic colour scale; (b) and (d) show each row normalised to its true class.

TABLE VI

MODEL COMPARISON UNDER A COMMON WEIGHT AND TOP-FRACTION CONVENTION (FIXED 13%). THE CONSTANT PREDICTOR HAS TIED SCORES (AUC 0.5) AND SELECTS NO EVENTS (AMS 0) BY CONSTRUCTION.

Model	Val. wtd. AUC	Val. raw AMS	Hold. wtd. AUC	Hold. raw AMS	Hold. $\times 5$ AMS ($b_{\mathrm{reg}} = 10$)
ANN ensemble, 5 seeds	0.937088	1.6043	0.933462	1.5552	3.5007
Histogram gradient boosting (tuned)	0.934768	1.5557	0.931935	1.5161	3.4092
Shallow NumPy ANN (16 units)	0.933279	1.5126	0.929563	1.4952	3.3631
Logistic regression (tuned, train-fitted)	0.883096	0.6915	0.880832	0.7072	1.5851
Constant predictor	0.500000	0.0000	0.500000	0.0000	0.0000

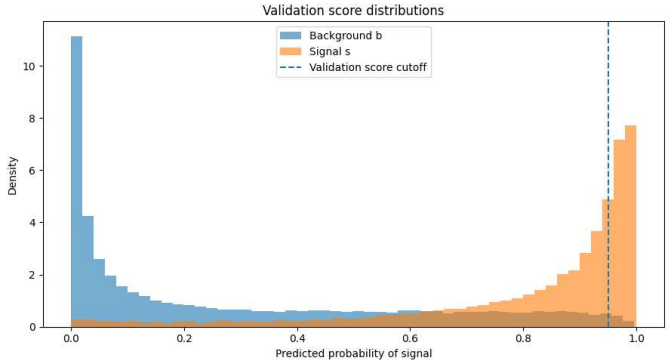


Fig. 9. Validation score distributions of the ensemble for background (b) and signal (s). Each class is normalised to unit area; the dashed line is the validation score cutoff of the selected fraction.

depend on the operating point.

Versus histogram gradient boosting. The ensemble's AUC is higher by 0.0023 on validation and 0.0015 on the holdout, and its holdout raw AMS is higher by 0.040, but the paired AMS interval is $[-0.032, 0.124]$ and contains zero. A single ANN member (mean AUC 0.9361 validation, 0.9325 holdout) exceeds boosting by about 0.0014 and 0.0006, the latter only about 1.5 member standard deviations; no paired AUC interval was computed. This comparison therefore does not establish that the ANN improves on boosted trees; an ensemble of tuned networks is being compared with a single, briefly tuned tree model.

Cost. The five-member ensemble has 31,045 parameters and took 335 s to fit in total (inference on the 50,000 validation events: 0.29 s). The tuned logistic model has 64 coefficients (search: 11.9 s; inference 0.006 s) and the tuned boosting model 6,100 tree nodes (search: 44.4 s; inference 0.63 s). Timings are single wall-clock measurements in the Colab runtime and are indicative only.

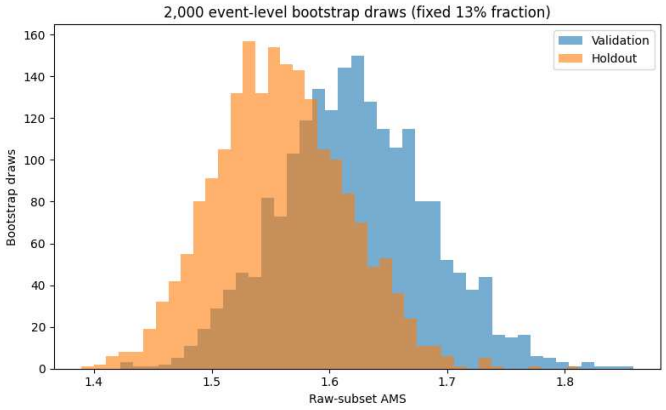


Fig. 10. Bootstrap distributions of the raw-subset AMS (2,000 event-level resamples) for the validation and holdout sets, with the selected event fraction held fixed.

D. Preprocessing and objective ablations (directional)

Table VII and Fig. 11 compare variants on the validation sample. These are *not* clean one-factor tests. Each preprocessing variant refits the training-only preprocessing, then trains one network (seed offset differs per variant) with the selected configuration for at most 35 epochs (patience 7); their best epochs (32–34) are close to that cap, whereas the five reference members were allowed up to 80 epochs (best epochs 31–69). The reference used below is therefore the mean of the five single members (AUC 0.936131 ± 0.000175, raw AMS 1.5796 ± 0.0390), not the ensemble. The unweighted-objective variant was repeated with five seeds under the same 35-epoch budget.

With that caveat: removing the logarithmic transform or all missingness and invalid-value flags lowers AUC by 0.0006 and 0.0009, respectively (several times the seed SD of the members, but each from one budget-limited run); replacing $\sin / \cos$ by raw azimuths or merging the two- and three-or-more-jet categories changes AUC by less than 0.0001. No preprocessing variant changes raw AMS by more than about 0.033, which is smaller than the seed-to-seed SD of 0.039 and far inside the bootstrap interval. The unweighted objective gives a lower AUC (by 0.0052, about four unweighted-seed SDs) and a slightly lower raw AMS at 13% (by 0.031, within the seed SD). A narrow low-fraction AMS optimum for the unweighted objective, which a single run could suggest, is not reproducible across seeds: its five seed-wise validation peaks lie at 11.0–15.5% (mean $13.4\% \pm 1.9\%$), and a 500-resample bootstrap of the peak location of the seed-averaged scores gives a mean of 13.1% (SD 2.5%). The weighted objective is therefore favoured on AUC, while its advantage in AMS is not established, and the weighted/unweighted comparison remains confounded with training budget. The one-hidden-layer network, trained for at most 50 epochs as a single run, scores lower than the two-layer members in AUC, but depth is not isolated from seed and budget effects.

E. Permutation feature importance

Permuting `DER_mass_vis` reduces validation weighted AUC from 0.937088 to 0.819043 (drop 0.118045). Other prominent variables are `DER_deltar_tau_lep` (0.082266), `PRI_tau_pt` (0.057100), `DER_mass_MMC` (0.050022) and `DER_mass_transverse_met_lep` (0.025224); the `DER_mass_MMC` missingness flag (0.011604) ranks seventh (Fig. 12). This is physically plausible: visible mass, angular separation, transverse momentum and reconstructed mass encode the tau-lepton final-state kinematics. The ranking is conditional on this model, representation, validation sample and metric. Correlated variables such as `DER_mass_MMC` and `DER_mass_vis` share information, so permuting one may understate the importance of the group. Permutation importance is a diagnostic, not a causal attribution.

X. Discussion

RQ1 (what matters). In the recorded runs the logarithmic transform and the missingness/invalid-value indicators are associated with small AUC losses when removed (0.0006 and 0.0009), whereas $\sin / \cos$ encoding and the split between two-jet and three-or-more-jet events show no measurable effect; no preprocessing change moves raw AMS outside seed noise. Because each variant is a single shorter-trained model, none of these gaps can be assigned to a single component with confidence. The permutation ranking shows that the signal is carried mainly by mass and tau-lepton kinematics, with the `DER_mass_MMC` missingness flag as the only indicator among the leading features.

RQ2 (conventions). The weighting and AMS conventions change the scale of AMS but not, by themselves, the ranking of selected sets: the scale-consistent convention is exactly $\sqrt{5}\times$ the raw value, and the fixed-regulator convention differs from it only through the regulator. The operating point and the primary selection metric matter more. Model selection by AUC and by AMS can disagree (configuration 1 has the highest raw AMS in the search; configuration 4 has the highest AUC), and AMS at a single fixed fraction is noisy. A global ranking metric and a single-point metric answer different questions, so both are reported.

RQ3 (what survives). Table VIII lists the claims and their status. The advantage over the train-fitted logistic baseline survives every check. The advantage over histogram gradient boosting does not survive the paired AMS test and is at most a small AUC edge. The progression from the historical logistic run to the networks is evidence of workflow evolution, not an isolated architectural effect, because the historical run differs in preprocessing, split and feature set.

XI. Limitations and Reproducibility

Limitations. The data are simulated; detector-calibration uncertainties (jet and tau energy scales and resolutions, missing-momentum modelling, identification and trigger efficiencies, luminosity, theory and background modelling) are

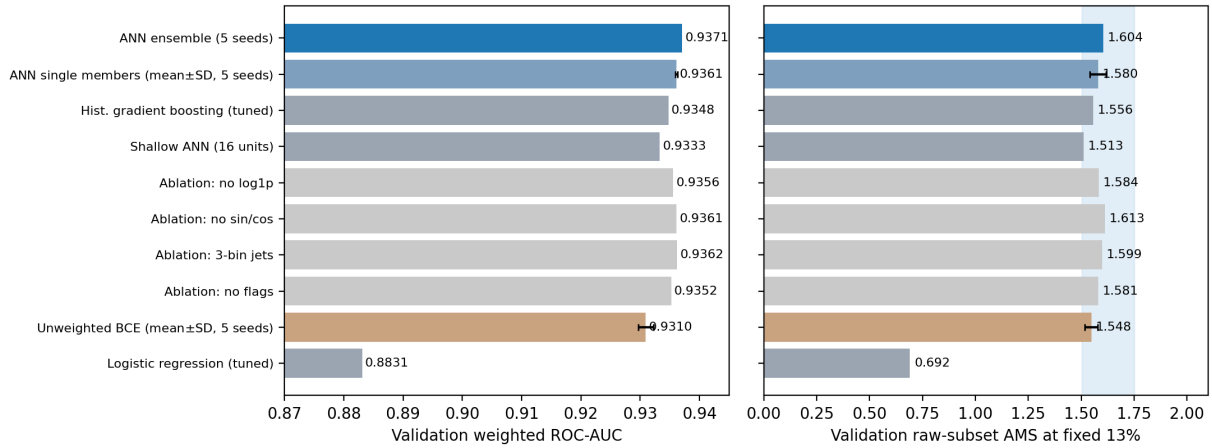


Fig. 11. Same-split validation comparison of models and ablations at the fixed 13% fraction (values from Tables VI and VII). Error bars on multi-seed rows are ± 1 sample SD. The shaded band in (b) is the 95% event-bootstrap interval of the ensemble's raw AMS; differences inside it should not be read as established.

TABLE VII

DIRECTIONAL ABLATIONS ON THE VALIDATION SAMPLE (FIXED 13% FRACTION). REFERENCE: MEAN OF FIVE SINGLE ANN MEMBERS (AUC 0.936131, RAW AMS 1.5796, SD 0.0390). VARIANTS ARE SINGLE 35-EPOCH MODELS UNLESS STATED; DIFFERENCES ARE BUDGET-CONFOUNDED.

Variant	Inputs	AUC	Raw AMS	ΔAUC	ΔAMS
No $\log(1+x)$	49	0.902378	1.069456	-0.0006	$+0.004$
No $\sin$ / $\cos$ (raw ϕ)	58	0.936179	1.553634	$+0.0000$	$+0.033$
Three jet bins (0, 1, 2+)	62	0.936429	1.621831	$+0.0001$	$+0.020$
No missing/invalid flags	38	0.935704	1.565085	-0.0009	$+0.002$
One hidden layer (16), 50 ep.	63	0.932962	1.488773	-0.0028	-0.067

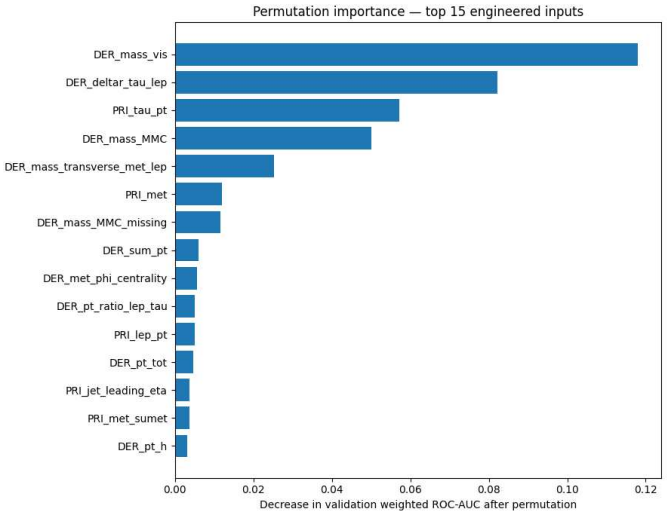


Fig. 12. Top validation permutation-importance features (drop in weighted ROC-AUC). The ranking reflects this fitted model and evaluation sample.

propagated in real ATLAS analyses through nuisance parameters in a likelihood fit [10], [11] and are absent here. The event-bootstrap and paired-bootstrap intervals describe evaluation-sample variability conditional on the fitted scores; they do not reflect retraining or selection variability. Seed-wise intervals rest on five seeds and cover only initialisation and data-order variability of the final configuration. Logistic regression, boosting, the shallow network and all ablations are single fits; the hyperparameter search used a 40,000-row training subset, short training and a single seed; and the ablations are confounded with training budget (35 versus up to 80 epochs). Gradient boosting was tuned over only four settings. A full uncertainty study would repeat the entire training and tuning pipeline over independent seeds and splits. The 14 invalid-value indicator columns are presumed constant in these data (Section VI). Finally, the same internal holdout was also scored in an earlier iteration of this pipeline with different preprocessing and a different selection rule; no choice in the present version was tuned against it, but it is not untouched in the strictest sense and should not be reused to choose between objectives, features or fractions. The 13% fraction likewise originates in an earlier validation sweep.

Reproducibility. The workflow fixes a global seed (42), records the five ensemble seeds (243–247) and their best epochs, fits preprocessing and class-normalisation factors on training data only, logs package versions (NumPy 2.4.6, pandas 2.2.3, scikit-learn 1.6.1, Matplotlib 3.10.0) and saves the feature list, transformation statistics, imputation values, log-transform rule, jet encoding and fixed fraction with the parameters of all five members in a compressed `.npz` bundle. Seed-wise validation and holdout metrics, model-comparison tables, per-event validation and holdout predictions (scores, labels,

TABLE VIII
CLAIMS LEDGER: WHAT THE RECORDED EVIDENCE DOES AND DOES NOT SUPPORT.

Claim	Evidence	Status
Analytical gradients are implemented correctly	Finite-difference check (1 hidden layer, unweighted, $\lambda = 0$), max relative error 6.552×10^{-10}; logistic loop and vectorised results agree	Established for the tested path; weighted, L2 and 2-layer paths untested
ANN ranks events much better than a tuned logistic baseline	Val. AUC 0.937 vs 0.883; holdout paired raw-AMS difference 0.851, interval [0.753, 0.960]	Supported (single split)
ANN improves on gradient boosting	AUC $+0.0023$ (val.), $+0.0015$ (holdout); paired holdout raw-AMS difference 0.040, interval $[-0.032, 0.124]$; 5-seed ensemble vs one briefly tuned tree model	Not established
Weighted objective is superior to unweighted	Higher AUC (0.9361 vs 0.9310) and slightly higher raw AMS (1.580 vs 1.548, within seed SD); 80- vs 35-epoch budgets	AUC favoured; AMS not established; budget-confounded
Log transform and missingness flags help	ΔAUC -0.0006 / -0.0009 when removed; ΔAMS within seed SD; single 35-epoch runs	Directional only
Sin/cos encoding and 4- vs 3-bin jets help	ΔAUC ≤ 0.0001, ΔAMS within seed SD	Not supported (no effect seen)
Depth helps	One-hidden-layer AUC -0.0028; single run, 50 vs up to 80 epochs	Directional only
Holdout performance is consistent with validation	AUC 0.9335 vs 0.9371; raw AMS 1.555 vs 1.604; bootstrap intervals overlap	Consistent; validation modestly optimistic
13% is the optimal operating fraction	Seed-wise validation peaks at 14.5–16.0%; bootstrap peak 12–18% (5–95%)	Not claimed; fixed convention
Output scores are calibrated probabilities	ECE 0.2045; Brier 0.1126	Not supported
Results constitute a Higgs discovery or physics measurement	Simulated data; no systematics or likelihood model	Not claimed

weights; including per-seed and unweighted-seed scores) are exported as CSV files. Models are implemented in NumPy [18]; the reference boosted model and logistic baseline use scikit-learn [8]. The run also wrote `submission.csv` with fields `EventId`, `RankOrder` and `Class` for all 550,000 test events; no AMS can be computed for it locally. Notebooks, input-file checksums, runtime versions and the generated submission should accompany the paper. *[Repository URL to be inserted before submission.]*

XII. CONCLUSION

This paper offers a reproducible, leakage-controlled educational framework for $H \to \tau\tau$ classification on the ATLAS open labelled data, not a new classifier. Its measurable findings are that, on one shared split, a from-scratch two-hidden-layer network ensemble (holdout weighted AUC 0.9335; raw AMS 1.56 with a 95% event-bootstrap interval of [1.45, 1.67]) separates signal from background far better than a tuned logistic baseline; that individual seeds are stable (AUC SD 0.0004, raw AMS SD 0.02 on the holdout); that the ensemble's margin over tuned histogram gradient boosting is not distinguishable from evaluation-sample noise (paired AMS difference 0.040, interval $[-0.032, 0.124]$); and that preprocessing and objective choices produce at most small, budget-confounded differences in AUC and none outside seed noise in AMS. The $\times 5$ AMS conventions rescale yields rather than adding evidence, and the claims ledger records what is and is not established. The most valuable next steps are to rerun the ablations with single-model references under identical training budgets and several seeds, to repeat the whole tuning pipeline over independent splits, to add a paired AUC comparison with boosted trees, and to evaluate calibrated outputs on independent calibration data. None of this replaces a profile-likelihood analysis with detector and theory nuisance parameters.

ACKNOWLEDGMENT

The author acknowledges the ATLAS Collaboration and the CERN Open Data Portal for releasing the simulated dataset and its documentation used in this educational study.

XIII. VIDEO EXPLANATION

https://youtu.be/TZR9yCzc3xU

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*, vol. 716, no. 1, pp. 1–29, 2012.

[2] S. Chatrchyan and others (CMS Collaboration), "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys. Lett. B*, vol. 716, no. 1, pp. 30–61, 2012.

[3] A. M. Sirunyan and others (CMS Collaboration), "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector," *Phys. Lett. B*, vol. 779, pp. 283–316, 2018.

[4] R. L. Workman and others (Particle Data Group), "Review of particle physics," *Prog. Theor. Exp. Phys.*, vol. 2022, no. 8, p. 083C01, 2022.

[5] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature Commun.*, vol. 5, p. 4308, 2014.

[6] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, "The Higgs boson machine learning challenge," in *Proc. NIPS 2014 Workshop on High-Energy Physics and Machine Learning, Proc. Mach. Learn. Res.*, vol. 42, 2015, pp. 19–55.

[7] ATLAS Collaboration, "Dataset from the ATLAS Higgs boson machine learning challenge 2014," CERN Open Data Portal, 2014, doi: 10.7483/OPENDATA.ATLAS.ZBP2.M5T8. [Online]. Available: https://opendata.cern.ch/record/328.

[8] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[9] ATLAS Collaboration, "The ATLAS detector," ATLAS Experiment, CERN, [Online]. Available: https://www.atlas.cern/Discover/Detector.

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

[11] ATLAS Collaboration, "Measurement of the CP properties of Higgs boson interactions with τ-leptons with the ATLAS detector," *Eur. Phys. J. C*, vol. 83, 2023.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015, arXiv:1412.6980.
[14] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall/CRC, 1993.
[15] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML), Proc. Mach. Learn. Res.*, vol. 70, 2017, pp. 1321–1330.
[16] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 1999, pp. 61–74.
[17] Scikit-learn developers, "Probability calibration," scikit-learn user guide, accessed Oct. 2026. [Online]. Available: https://scikit-learn.org/stable/modules/calibration.html.
[18] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.