AUA
American University of Armenia

ML4JETS
VIENNA ▪ 14–18 SEPTEMBER 2026

# A Machine Learning Study for Higgs Boson Detection

Signal discrimination in H -> ZZ* -> 4l with ATLAS Open Data

Mane Papoyan

Supervisor: Houry Keoshkerian

14-09-2026

# Introduction

- **Self-introduction**
  - **Mane Papoyan**, recently graduated with a bachelor's degree in the **Data Science** program at the **American University of Armenia**
  - This analysis is the result of my final-year Capstone project

*Presenting a Machine-learning analysis of the Higgs boson in the H→ZZ*→4ℓ decay channel with the ATLAS Open Data at 13TeV, corresponding to 36.6 fb-1*

- Event selection and the 4-lepton system reconstruction
- Physics-motivated feature engineering and selection of kinematic variables
- Comparison of seven classifiers trained on simulated signal and background events.
- Classifier optimization and signal significance evaluation using MC and Data-Driven background-estimation.
- Application to real ATLAS collision data and identification of the Higgs signal.
- Conclusion and outlook

# Why the 4-lepton Channel

**Advantages**

- Clean experimental signature with four identifiable charged leptons.
- Fully reconstructable final state with excellent four-lepton invariant-mass resolution.
- Clear Higgs mass peak near 125GeV.
- Well-understood background processes and rich kinematic information for machine-learning classification.

**Challenges**

- Extremely rare final state: ~ 30 signal events in 36.6fb−1.
- Significant class imbalance: ~1,200 background events.
- Irreducible $ZZ^{*} \to 4\ell$ background.
- Effective identification requires maximizing signal retention while suppressing background.

*The four-lepton channel provides a clean Higgs signature, while machine learning may improve signal–background discrimination and enhance sensitivity.*

# Data and Monte-Carlo Simulated events

**Data**

- ATLAS Open Data: proton-proton collisions at 13 TeV, 36.6 fb^-1 from 2015 - 2016

**Monte-Carlo (POWHEG+Pythia8, Sherpa)**

- **Signal**
  - $H \to ZZ^* \to 4\ell$, with mH=125GeV.

- **Background**
  - **Irreducible:** ZZ* -> 4l
  - **Reducible:** Z+jets, ttbar, ttV, and triboson, where extra leptons come from decays or fakes.

*Simulated events are normalized to the integrated luminosity and corrected for detector effects, enabling direct comparison with observed data.*

# Data and Monte-Carlo Simulated events

Why "irreducible" versus "reducible"

- $ZZ^* \to 4\ell$ produces four genuine prompt leptons: the same visible final state as the signal, so no lepton-quality requirement can remove it. Only kinematics separate it.
- Reducible backgrounds contribute through non-prompt or misidentified leptons, so identification, isolation, and pT requirements suppress them directly.

Technical setup

- Samples distributed as ROOT files, the columnar format standard in high-energy physics.
- Accessed through the atlasopenmagic Python package, delivered as awkward arrays for direct Python analysis.

# Event Selection and Yields

**Applied in order**

1. exactly4lep skim: exactly four reconstructed leptons.
2. Lepton reconstruction, identification, and isolation.
3. Same-flavour, opposite-sign pairing only: 4e, 4μ, or 2e2μ.
4. Total charge of the four leptons = 0.
5. pT staircase (ordered): $\ell_1 > 20$, $\ell_2 > 15$, $\ell_3 > 10$, $\ell_4 > 7$ GeV.

| Lepton | Requirement |
|---|---|
| $l_1$ | $p_T > 20$ GeV |
| $l_2$ | $p_T > 15$ GeV |
| $l_3$ | $p_T > 10$ GeV |
| $l_4$ | $p_T > 7$ GeV |

**Why the staircase**

- pT is the momentum transverse to the beam, well measured at a hadron collider.
- A hard leading lepton plus softer companions suppresses Z+jets and $t\bar{t}$, whose extra leptons are typically soft.

**Unweighted yields after preselection**

- 226,168 simulated signal events
- 12,436 simulated background events
- 1,279 observed events in data

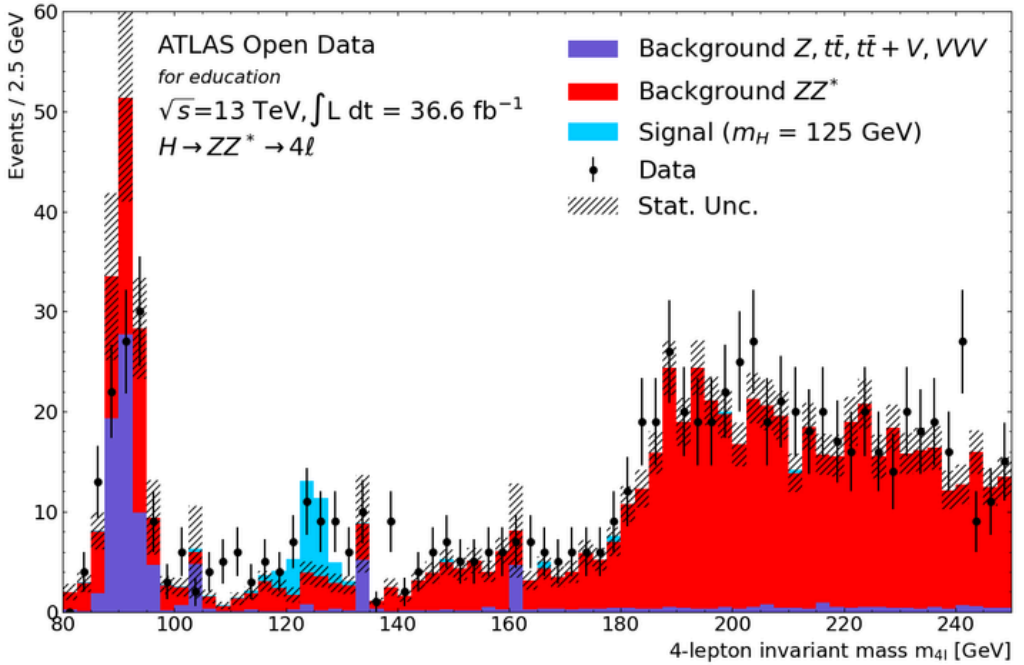
# $Z_1$ / $Z_2$ reconstruction and $m_4\ell$

**Pairing**

- Four leptons give at most three pairings: $(\ell_1\ell_2)(\ell_3\ell_4)$, $(\ell_1\ell_3)(\ell_2\ell_4)$, $(\ell_1\ell_4)(\ell_2\ell_3)$.
- Only same-flavour, opposite-sign pairs are valid Z candidates.
- The pair closest to mZ = 91.2 GeV is $Z_1$ (on-shell); the other is $Z_2$ (off-shell). This is the ATLAS convention.

**The resulting observable**

- $Z_1 + Z_2$ is the Higgs candidate; $m_4\ell$ is the main kinematic variable.




*Pairing the four leptons into two Z candidates gives the event its physical structure, and their combination gives the analysis its key observable.*

# Event weighting and expected yields

**Cross-section weight**

- Each MC event is scaled by $L \times \sigma \times \eta \times k \times w\_MC$, then normalised by the sum of generator weights.
- $L = 36.6\ fb^{-1}$; $\sigma$, $\eta$, $k$, and $w\_MC$ are stored per event and differ by process.

**Detector scale factors**

- Multiply by pile-up, electron and muon reconstruction/ID, and lepton-trigger scale factors.
- The total weight is used for all histograms and yields.

*Per-event Monte Carlo weights put simulation on the same luminosity footing as data, and they reveal the true 1:40 signal-to-background regime.*

| Process | Raw Yield | Weighted Yield |
|---|---|---|
| $H \to ZZ^* \to 4\ell$ (Signal) | 226,168 | $30.43 \pm 1.33$ |
| $ZZ^* \to 4\ell$ (Irreducible Bkg) | 6,364 | $1067.08 \pm 18.15$ |
| Reducible Backgrounds | 6,033 | $127.61 \pm 16.32$ |
| Total Background | 12,436 | $1194.69 \pm 24.41$ |
| Data | | $1{,}279 \pm 35.8$ |

# Feature engineering: 31 kinematic variables

**Seven groups**

1. Z properties: $m(Z_1)$, $m(Z_2)$, $pT(Z_1)$, $pT(Z_2)$
2. Four-lepton system: $pT(4\ell)$
3. After Z assignment: pT and η of each lepton (8 features)
4. Helicity angles: $\theta_1$, $\theta_2$, $\Theta$, $\Phi$, $\Phi_1$ (5 features)
5. Angular separations: $\Delta\phi$ within each Z, plus four cross-pair $\Delta R$
6. Scalar pT sums within each Z
7. Event-level: N_jet, MET, $\phi$_MET

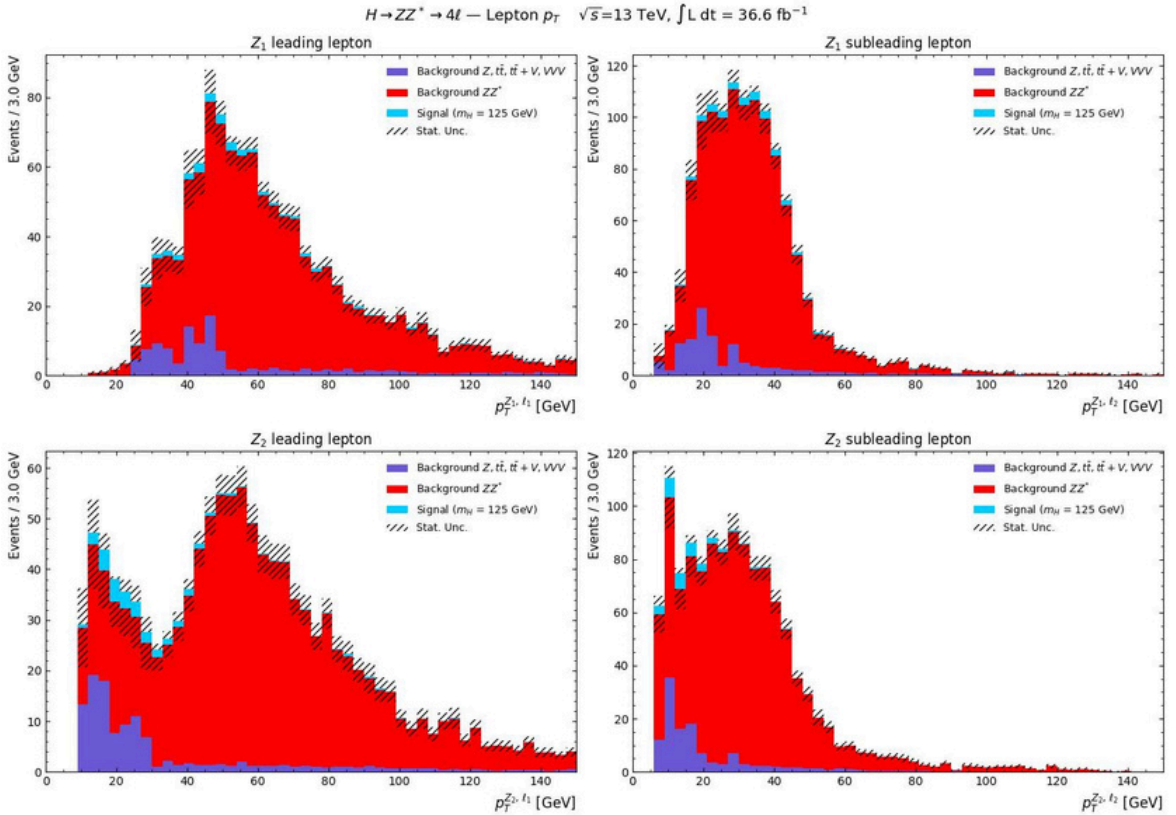
**Helicity angles**

- Boost leptons into the Z rest frames → $\theta_1$, $\theta_2$
- Boost the $4\ell$ system into the Higgs rest frame → $\Theta$, $\Phi$, $\Phi_1$

*Four lepton four-vectors become 31 physics observables: kinematics plus angular structure.*

# Reading the feature distributions

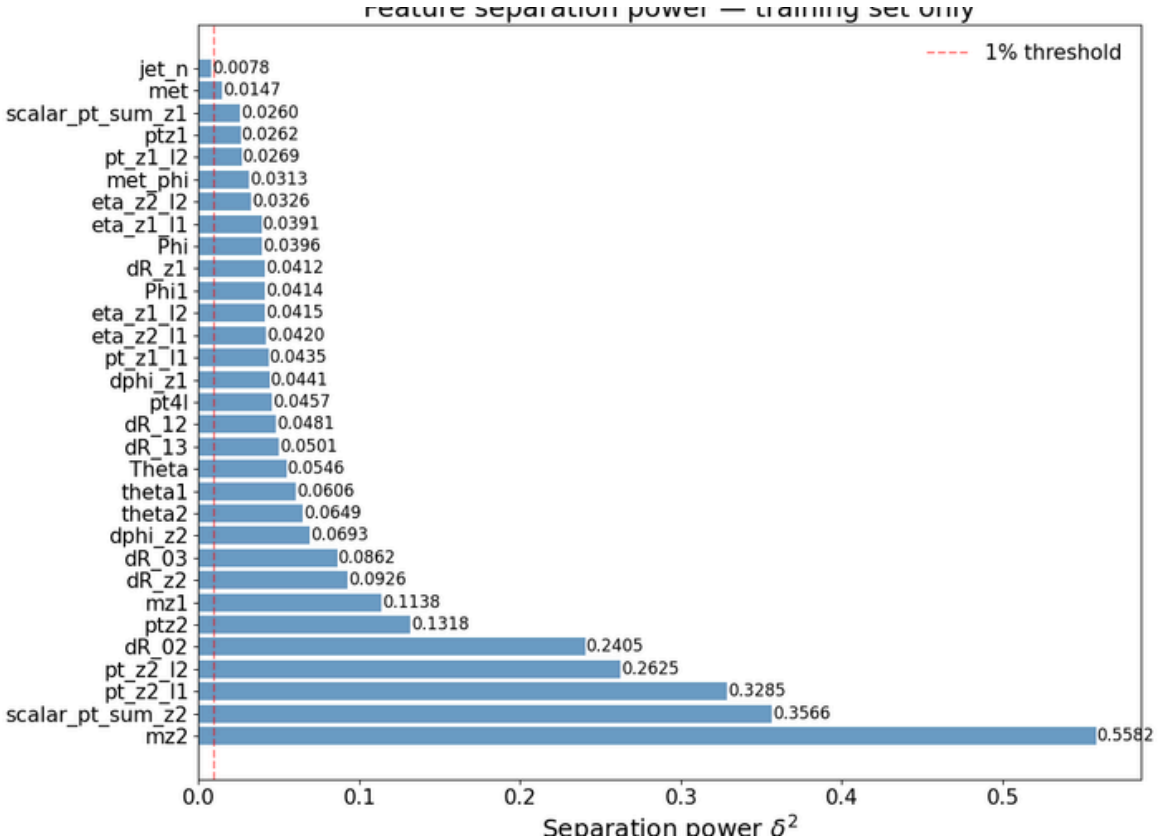
- Histogram all 31 features for signal vs background.
- This checks that each variable behaves as expected and shows why it can discriminate.
- Example: lepton pT shapes differ because a 125 GeV resonance is not the same as continuum ZZ*.
- Those shape differences are what the next slide's ranking quantifies.




*Before any model is trained, comparing signal and background histograms shows where the discriminating information physically lives.*

# Ranking features by separation power

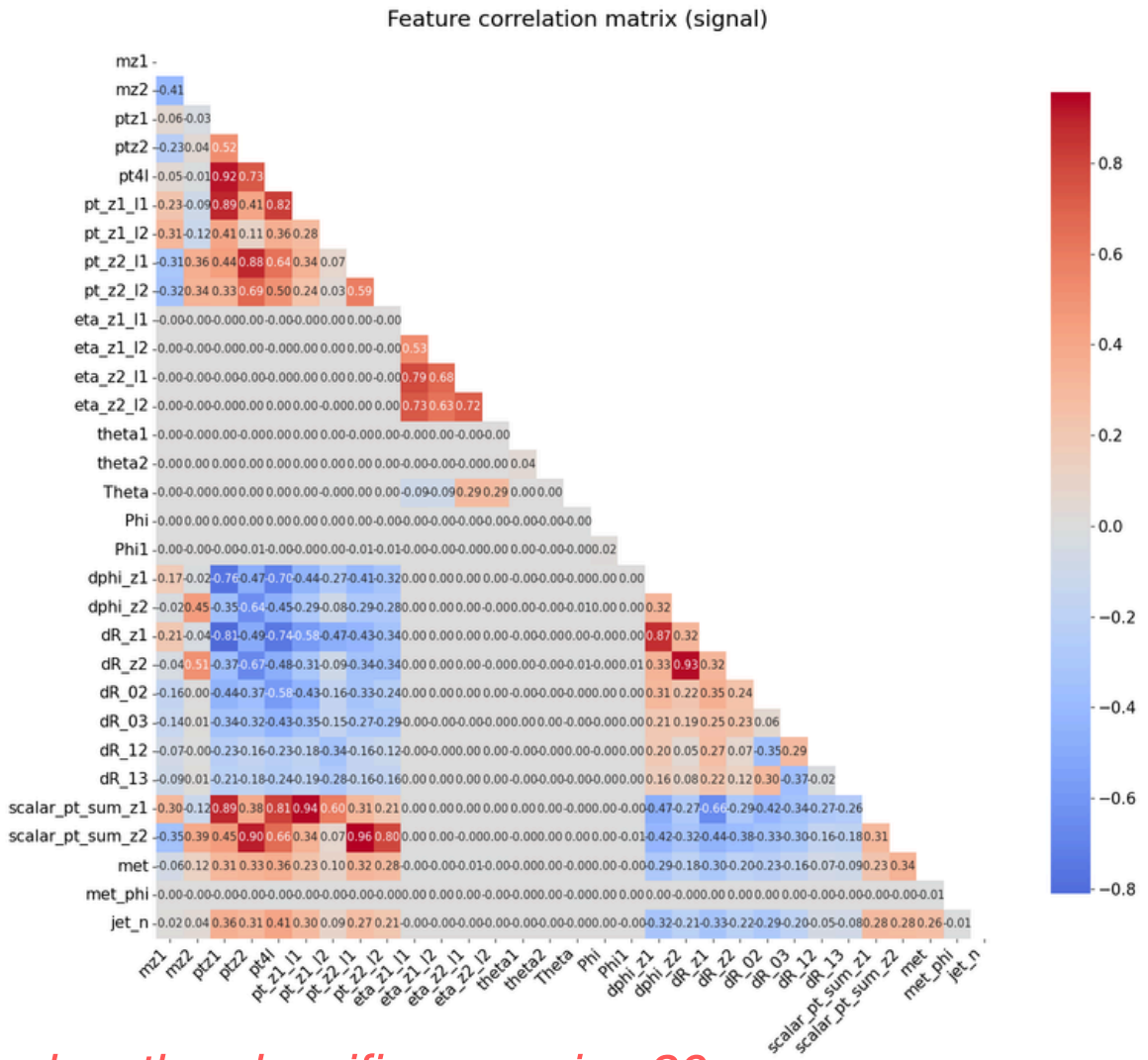
- $\delta^2$ is computed from normalised signal and background histograms: squared bin-by-bin difference relative to their sum (50 bins per feature).
- 0 = identical shapes; 1 = fully disjoint.
- Computed on the training set only, so feature choice cannot leak from evaluation events.
- Ranking is cross-checked, not used as the sole selection criterion.




*Each feature's discriminating power is quantified bin-by-bin, so feature selection rests on a measured quantity rather than intuition.*

# Pruning correlations: 31 → 20 inputs

- Identify redundant variables using the feature-correlation matrix.
- For pairs with $|\rho| \geq 0.7$, retain the variable with greater separation power.
- Reduce 31 variables to 20 ML input features.
- Exclude m4ℓ to avoid biasing the Higgs signal extraction.




*Highly correlated features are removed so the classifiers receive 20 non-redundant inputs.*

# ML Training protocol: seven classifiers, one framework

**The seven classifiers**

- XGBoost, LightGBM, Random Forest
- Multilayer perceptron (MLP), implemented in Keras
- Logistic Regression
- Quadratic Discriminant Analysis (QDA)
- Gaussian Naive Bayes (GaussianNB)

**Three model families:** boosted/bagged trees, a neural network, and parametric baselines. Baselines are there so any gain from the complex models is shown, not assumed.

**5-fold stratified CV (out-of-fold scores)**

- Split MC into five folds, preserving the signal–background mix.
- Train on four folds, score the held-out fold; every event gets one unseen score.
- Metrics, threshold scans, and yields all use these scores

*A common training and validation framework enables a consistent comparison of all seven classifiers.*

# Class imbalance, weights, and hyperparameters

**The imbalance problem**

- Summed signal weights ⋘ summed background weights.
- Uncorrected, a model learns “everything is background.”

**How it is handled**

- Boosted trees: scale the signal class by Σw_bkg / Σw_sig.
- Random Forest, MLP, Logistic Regression: class-balanced sample weights.
- QDA: weighted bootstrap of each training fold.

Luminosity weights are kept for evaluation. Balancing affects training only.

# Class imbalance, weights, and hyperparameters

**Tuned by inner cross-validated grid search**

- Boosted trees: 5-fold inner CV over maximum tree depth $\in \{3, 4, 5\}$ and minimum child weight $\in \{1, 5, 10\}$.
- Logistic Regression: 6-point inner scan over inverse regularisation strength $C \in \{0.001, 0.01, 0.1, 1, 10, 100\}$, with $C = 100$ selected.

**Fixed by design**

- MLP: three hidden layers of 128, 64, 32 units with ReLU activations, sigmoid output, Adam optimiser at learning rate $10^{-3}$, binary cross-entropy loss, up to 100 epochs, batches of 256, early stopping on a validation-loss plateau with patience 10.
- Random Forest: 500 trees, maximum depth 5, minimum 5 samples per leaf.

*Training is adapted to a rare-signal problem without distorting the physical event yields used for evaluation.*

# Robustness of the training setup

**Why k = 5**

- Repeated the analysis for k ∈ {3, 5, 7, 10}.
- Beyond five folds, fold-to-fold AUC scatter grows with no gain in mean AUC or downstream results.

**Seed stability**

- Repeated 5-fold training with 15 random seeds (folds and initialisation).
- AUC varies by < 0.005; downstream results by ~3%. Not an artefact of one split.

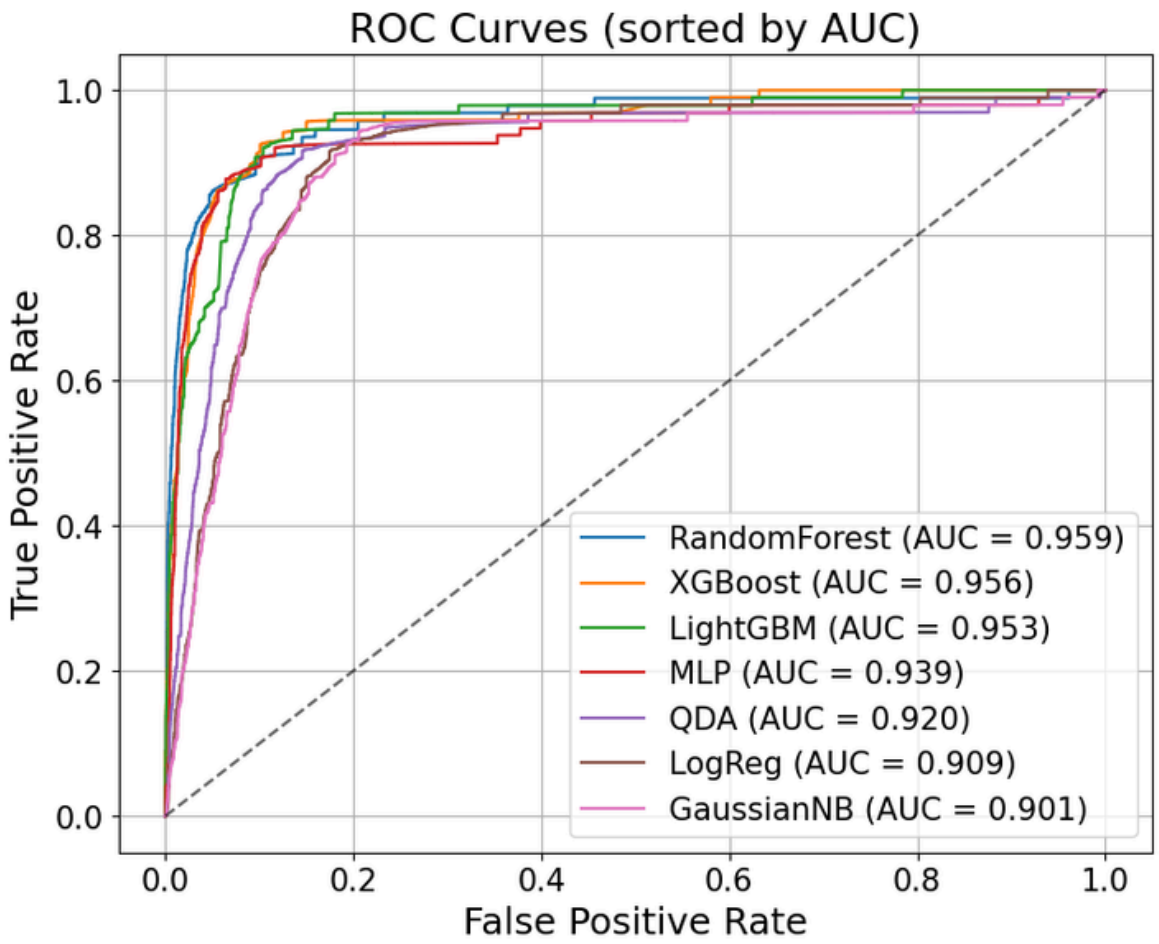
**AUC uncertainty**

- Stratified bootstrap of the out-of-fold scores (1,000 resamples, class mix preserved).
- Gives 95% CIs on the next slide, so models are compared with error bars.

*The choice of five folds, the random seed, and the finite sample size are each shown not to drive the results.*

# Classifier Performance: ROC Curves and AUC

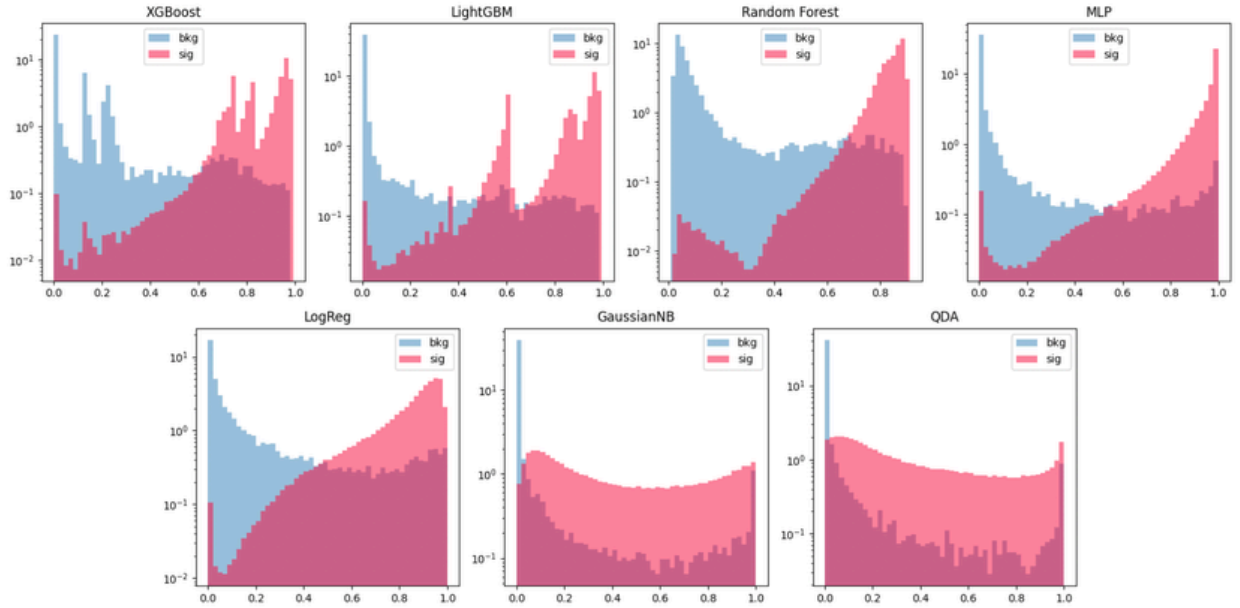
- Signal–background discrimination evaluated using ROC curves and out-of-fold AUC.
- AUC uncertainties estimated using 1,000 bootstrap resamples.
- XGBoost, LightGBM, and Random Forest achieve the highest AUC values of 0.956, 0.959, and 0.953, respectively
- MLP, QDA, Logistic Regression, and Gaussian Naive Bayes achieve AUC values of 0.946, 0.920, 0.909, and 0.901, respectively.




*XGBoost, LightGBM, and Random Forest show comparable discrimination performance, with overlapping confidence intervals.*

# Signal Efficiency and Background Rejection

- Evaluate classifiers on Monte Carlo simulation at a common score threshold of 0.65.
- XGBoost retains 87.65% of signal events while rejecting 92.71% of background.
- LightGBM retains 70.97% of signal events while rejecting 94.96% of background.
- QDA and Gaussian Naive Bayes achieve high accuracy but retain only 19.36% and 26.98% of signal events.
- Exclude QDA and Gaussian Naive Bayes from further analysis due to insufficient signal efficiency.




| Classifier | Accuracy | Signal Eff. | Bkg Rejection |
|---|---|---|---|
| XGBoost | 0.926 | 87.65% | 92.71% |
| LightGBM | 0.944 | 70.97% | 94.96% |
| Random Forest | 0.909 | 88.16% | 90.92% |
| MLP | 0.943 | 84.39% | 94.58% |
| Logistic Regression | 0.881 | 78.68% | 88.31% |
| QDA | 0.970 | 19.36% | 98.95% |
| GaussianNB | 0.955 | 26.98% | 97.25% |

*High accuracy can be misleading: effective classifiers must retain signal while rejecting background.*

# Background Estimation

- Apply the fixed classifier threshold and select events in the signal region $110 \leq m_{4\ell} < 135\,\text{GeV}$.
- **Monte Carlo estimate:**
  - B is taken directly from the weighted simulated background surviving both the classifier cut and the mass window.
  - These events are exactly the classifier's false positives.
  - Depends on simulation modelling the background correctly.
- **Data-driven estimate:**
  - B is estimated entirely from data, using no simulation at all.
  - Real events passing the same classifier cut are counted in the sidebands $90 \leq m_{4\ell} < 105$ GeV and $140 \leq m_{4\ell} < 155$ GeV.
  - The count is scaled by the ratio of signal-region width to total sideband width, assuming the background varies smoothly across the region.
  - Depends on that smoothness assumption, not on simulation.

*Consistent background estimates support a reliable calculation of the observed signal significance.*

# Background Estimation

**Outcome across the five classifiers**

- XGBoost: 17.96 ± 2.40 background events from simulation versus 16.67 ± 3.73 from the sidebands, consistent within uncertainties.
- LightGBM: also consistent across both methods.
- MLP and Logistic Regression show large shifts between the two estimates, indicating sensitivity to background-modelling assumptions and therefore a lack of robustness.
- Random Forest is consistent between methods but gives a weaker overall result.

| Classifier | MC prediction $B$ | Sideband $B$ |
|---|---|---|
| XGBoost | 17.96 | 16.67 |
| LightGBM | 15.13 | 14.17 |
| Random Forest | 21.66 | 23.33 |
| MLP | 28.49 | 17.50 |
| Logistic Regression | 19.91 | 35.83 |

# Signal Significance on ATLAS Data

Apply the five retained classifiers to the 1,279 observed events.

Keep events with score > 0.65 and 110 ≤ $m_4\ell$ < 135 GeV.

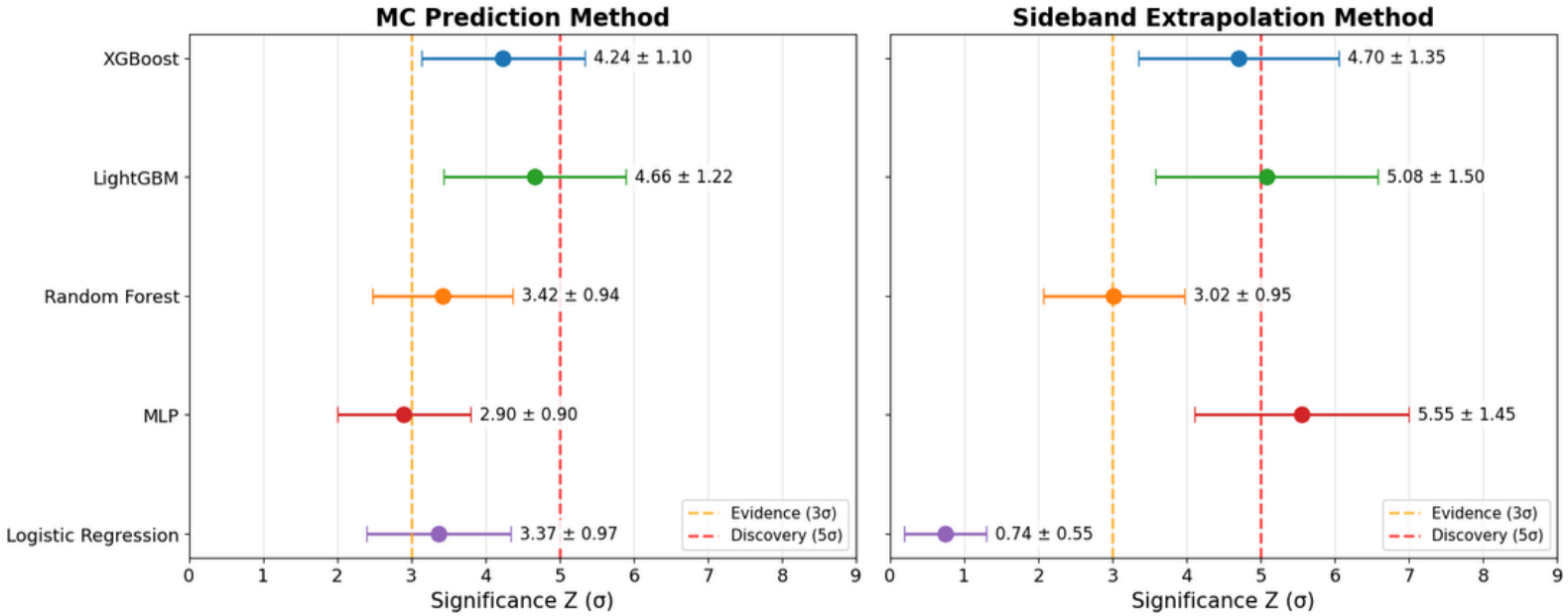
$$Z_{MC} = \frac{S}{\sqrt{B+\kappa^2 B^2}}, \qquad \kappa = 0.30$$

**Count-based significance**

- S = N_obs – B, with B from each of the two background estimates.
- Local counting estimate (one fixed window, no look-elsewhere correction):

XGBoost: 4.24±1.10σ (Monte Carlo) and 4.70±1.35σ (sidebands).

LightGBM: 4.66±1.22σ (Monte Carlo) and 5.08±1.50σ (sidebands).




*XGBoost and LightGBM improve the significance relative to the no-ML selection and remain consistent across both background-estimation methods.*

# Conclusion and Outlook

- Machine-learning classifiers improve signal–background discrimination in the $H \to ZZ^* \to 4\ell$ channel.
- XGBoost and LightGBM provide consistent evidence for the Higgs signal in  ATLAS data.
- Results are validated using two independent background-estimation methods.
- Signal significance improves compared with the analysis without ML selection.

**Outlook**

- Implement a profile-likelihood-based significance estimation.
- Develop more detailed background selection and identification methods.
- Model individual systematic uncertainties more accurately.
- Investigate simulation-to-data differences and extend the analysis to larger datasets.

# Thank You

# Backup Slides

# ML Classifier Hyperparameters

| Classifier | Settings |
|---|---|
| XGBoost | $n_{\text{est}} = 500$, max_depth and min_child_weight tuned (see Table XI-B), learning_rate $= 0.05$, subsample $= 0.8$, colsample_bytree $= 0.8$, $\gamma = 0$, $\lambda = 1$, $\alpha = 0$, `tree_method`=hist, early_stopping_rounds $= 20$ |
| LightGBM | $n_{\text{est}} = 500$, max_depth $= 4$, num_leaves $= 15$, learning_rate $= 0.05$, subsample $= 0.8$, subsample_freq $= 1$, colsample_bytree $= 0.8$, min_child_samples $= 20$, $\lambda = 1$, $\alpha = 0$, early_stopping_rounds $= 20$ |
| Random Forest | $n_{\text{est}} = 300$, max_depth $= 5$, min_samples_leaf $= 5$, bootstrap $=$ True |
| MLP (Keras) | layers $= [128, 64, 32]$ + 1, activations: ReLU/ReLU/ReLU/sigmoid, optimizer = Adam, learning_rate $= 10^{-3}$, loss = binary cross-entropy, epochs (max) $= 100$, batch_size $= 256$, early_stopping patience $= 10$ on val_loss |
| Logistic Reg. | solver = lbfgs, max_iter $= 2000$, $C$ tuned over $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$, inner 3-fold CV used to select $C$, best $C = 100$ |
| QDA | reg_param $= 0.1$, weighted bootstrap resampling per fold |
| GaussianNB | `scikit-learn` defaults; native `sample_weight` used |

# 8. From a score to a countable excess

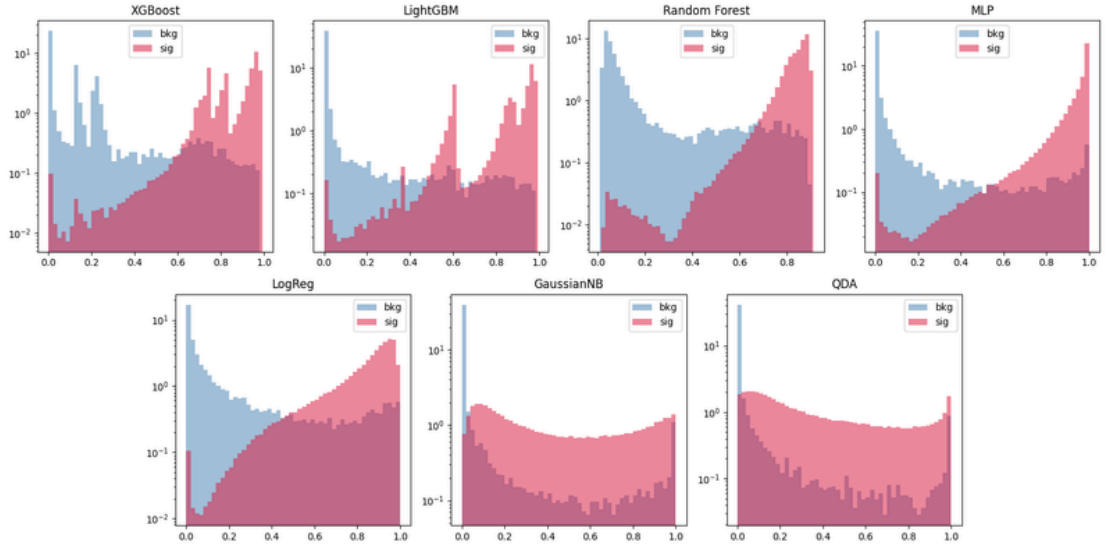
**One frozen threshold turns the score into a counting experiment: 2.3 to 3.6 sigma expected.**

**Scan on simulation only.** We take score > 0.65, chosen on the plateau rather than at the peak, so the result does not sit on a statistical spike.

**Then freeze it,** before a single real event is scored.

| Selection (simulation) | S | B | Expected Z |
|---|---|---|---|
| Mass window only | 25.8 | 31.9 | 2.32 +/- 0.36 |
| **+ score > 0.65** | **24.9** | **18.0** | **3.63 +/- 0.40** |

**The cut keeps 96% of the signal and removes 44% of the background.**




**Every Z carries an error bar,** propagated from the uncertainties on S and B.

**Local significance.** One fixed window, no look-elsewhere correction, counting rather than a likelihood fit.