Particle Physics in 1D: Classification Accuracy for Signal and Background Events under Gaussian Noise

Manoli Pratikakis

Physics 2900

Department of Physics

David Marsh

Dr. Kisa Ranasinghe

December 10 2025

Abstract

High-energy physics experiments at the Large Hadron Collider generate millions of particle-collision events, where only a tiny fraction correspond to rare or interesting physical signatures. Machine learning classifiers are therefore critical for distinguishing genuine signal events from abundant background noise in these datasets. However, real detectors introduce measurement uncertainty, commonly modeled as Gaussian smearing, which can degrade classification performance. This study investigates how a gradient-boosted algorithm (XGBoost) responds to increasing levels of simulated detector noise in a simplified, one-dimensional toy model representing LHC collision data.

1 Introduction

Physicists at CERN record millions of collisions every time they conduct a test and to detect those collisions, they use high quality sensors that can give them 30 40 measured variables relative to each collision. This means the data accumulated per collision is 30 40 times the number collisions that took place. That is a lot of data, and one of the most important aspects of machine learning algorithms in the CERN environment is being able to discern what collision out of the millions a physicist wants to discover; this is called classification. These classifications are signified as *signal* and *background*, and their use in a machine learning algorithm context will guide and contextualize this study. These datasets generated by the Large Hadron Collider (LHC) are huge and require computational rigor to analyze and interpret these datasets, machine learning algorithms. Machine learning refers to AI or a subcategory of AI which simply refers to not having to explicitly program an AI to complete, interpret data, or generate diagrams. This is the practicality behind the algorithms and the sensors and how that relates to data collections and analyzation. However, the most important aspect of all of this is how do scientists classify a collision as worth studying, or not? They use classification algorithms which can signify major events that may be worth studying; such events may include the development of a Higgs Boson. However, due to noise from detector

resolution this may cause the algorithm to interpret data differently, leading to the research question:

> *To what extent can gradient-boosted algorithms maintain signal–background classification accuracy when Gaussian noise is introduced into simulated LHC collision data?*

A more recent and more important discovery relative to HEP and the LHC as a whole would be the discovery of the Higgs Boson particle in 2012 by Peter Higgs and François Englert in which the Higgs Boson particle is what supported the idea behind a Higgs field which is fundamental into understanding how mass comes into existence (Atlas, 2012; Giasemis, 2025; Guardian, 2025). Now that AI has become much more rampant over the past few years, physicists at CERN have begun to utilize AI to better understand and interpret the data recorded in the LHC. The new head director at CERN is now talking about AI and the new and improved advanced methods of finding and interpreting CERN data. He says that there will be a time soon where the LHC and the AI and tools available will be able to record the Higgs Boson particle coupling with a second Higgs Boson particle which would cause a surge in the scientific community because Higgs particles have only ever been recorded to exist by themselves, so to have two exist simultaneous would be groundbreaking (Guardian, 2025). So, after the Higgs Boson and general advancements made by physicists around the world, this is the stage we are at now, where AI is becoming increasingly accurate, useful, and efficient to use. For interpretation and analyzation of data that help scientists arrive at conclusions even faster.

2 Key Concepts

2.1 The Large Hadron Collider (LHC)

The LHC is the biggest particle accelerator in the world which takes protons and moves them around in multiple circles slowly speeding them up to the speed of light then finally smashing them

together in the middle of a sensor environment. All of the data generated by collisions such as mass, energy, momentum etc.. is data that scientists at CERN are analyzing every day.

2.2 Machine Learning in High-Energy Physics

As previously mentioned, Machine Learning in a HEP context is the usage of substantial amounts of data to be able to interpret trends and optimize algorithms without being needed to explicitly code. Although there are several Machine Learning algorithms, one in widely used one in physics is gradient boosted decision trees in which the algorithm sorts an output using certain parameters given to the algorithm. Other algorithms include deep learning, neural networks, and polynomial regression which is basically how physicists are able to regress nonlinearly, which is extremely useful when dealing with such a large amount of dynamic data (Giasemis, 2025). However in this study, the algorithm being used will be XGBoost which is a gradient boosted algorithm that simply builds a powerful predictive model that attempts to sort data accordingly (based on given parameters) that it is being trained on. After it builds that model using the training data it can then be used to test on real data for classifying outputs and thereafter analyzing those outputs.

2.3 Detector Resolution

Detector resolution is an important aspect and limitation of the collected data for CERN scientists as the data collected from the sensors at CERN is not 100% accurate. Detector resolution is inherent to all sensors, and they are the true comparison of the actual data versus the error that the machine can illustrate through fluctuations in data (Bortoletto, D). Data fluctuations typically have a conceptually general mathematical representation as follows:

$$x_{\text{measured}} = x_{\text{true}} + \mathcal{N}(0, \sigma).$$

In this study, detector resolution is represented entirely by Gaussian smearing (σ) applied to all

data points.

3 Hypothesis

As Gaussian detector noise increases, the classification accuracy of the gradient-boosted algorithm will decrease; the magnitude of the decrease will depend on the noise standard deviation.

4 Significance Statement

Machine Learning is a crucial and growing tool used for analyzing hadron collider data. This is exemplified in an article that goes over some information Professor Mark Thompson says about AI revolutionizing LHC data analyzation: "machine learning is paving the way for advances in particle physics." He also mentions that Machine Learning will aid in uncovering new information about the Higgs boson and the origin of mass which are both insurmountably important (Guardian, 2025). Although this research topic does not aim to uncover information about the Higg's boson, it could help classify collisions in which the Higgs bosons appears. This study would be significant by illustrating a developed simulation toy that generate random gaussian data and illustrate how adding noise on top of that data the algorithm would respond. This could be useful for physicists who decide to develop new algorithms as they may decide to consider realistic simulation which may need detector resolution information.

5 Objectives

- Generate 1D simulated signal and background distributions using Gaussian models.
- Apply Gaussian noise with adjustable standard deviation.
- Train an XGBoost classifier on noiseless data and evaluate it on noisy data.

- Measure changes in classification accuracy as noise increases.

6 Goals

- Replicate how detector resolution impacts event classification.
- Assess robustness of gradient-boosted algorithms under Gaussian smearing.
- Create a simplified toy model that captures essential ideas of LHC classification.

7 Aim

To explore how Gaussian noise affects the ability of gradient-boosted algorithms to classify particle collisions effectively.

8 Research Design

This study uses computational modeling to simulate one-dimensional LHC-like data and applies Gaussian noise to approximate detector resolution. The classifier used is XGBoost, trained on true (noise-free) data and tested on various noise conditions.

9 Variables

Independent variable: Magnitude of Gaussian noise σ.

Dependent variable: Classification accuracy (ROC AUC, accuracy, Signal-Background efficiency).

Controlled variables: Model type, data generation parameters, train/test split.

10 Materials and Resources

- Python, XGBoost
- UNIX terminal, Windows Subsystem for Linux
- Simulated 1D Gaussian dataset
- Noise generated via `np.random.normal()`
- Laptop CPU

11 Procedure

1. Generate clean noise-free data.
2. Sample $N = 2500$ points for both signal and background.
3. Use means μ_S, μ_B with labels 1 (signal) and 0 (background).
4. Apply event weights to reflect rarity: $w_S = 10$, $w_B = 100$.
5. Select noise levels such as $\sigma = 0,\ 0.001,\ 0.005,\ 0.01$, etc.
6. Apply Gaussian noise:

$$x_{\text{noisy}} = x_{\text{true}} + \mathcal{N}(0, \sigma).$$

7. Train the XGBoost classifier.
8. Evaluate accuracy, ROC curves, and confusion matrices.

12 Data Collection and Analysis

Accuracy measurements, ROC AUC, and confusion matrix outputs are recorded for each noise level. Data is visualized using LateX or Python, including plots of ROC curves illustrating classifier performance degradation.

13 Results

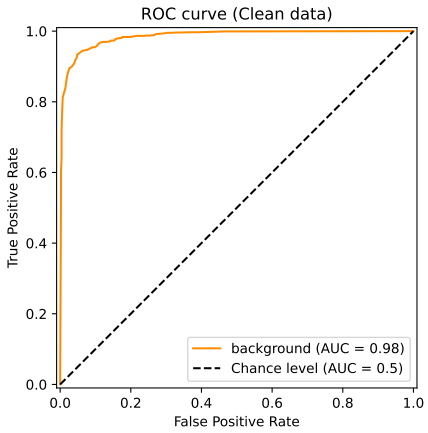

Figure 1: One run of test: Noise $\sigma = 0$

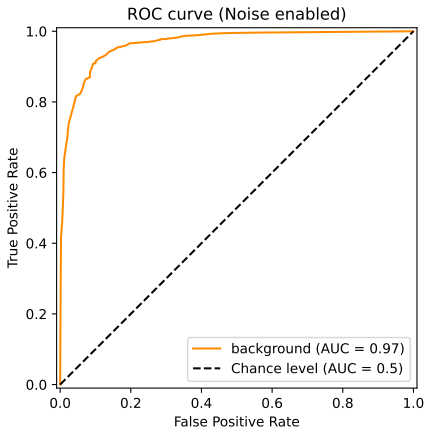

Figure 2: One run of test: Noise $\sigma = 0.5$

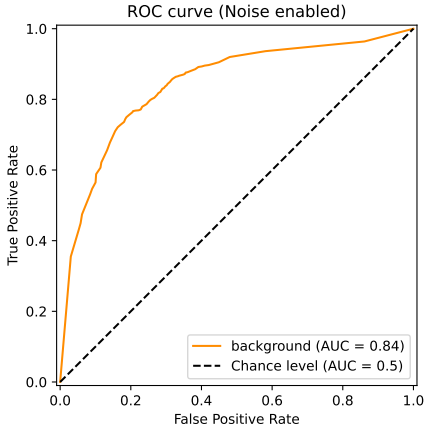

Figure 3: One run of test: Noise $\sigma = 1.5$

Noise σ	Runs	Avg Accuracy (%)	Avg Sig Eff	Avg Bkg Eff	Avg Final Sig
0	5	87.79	0.771	0.0112	7.20
0.05	5	88.04	0.775	0.0103	7.17
0.1	5	88.11	0.780	0.0167	7.21
0.5	5	86.89	0.757	0.0307	4.84
1.0	5	79.35	0.673	0.0709	2.60
1.5	5	75.28	0.644	0.1326	1.77

Table 1: Performance of XGBoost Classifier vs Detector Resolution Noise σ. Metrics averaged over repeated runs. Clean (noise-free) data included for comparison.

The performance of the XGBoost classifier was evaluated on datasets with varying levels of detector noise, quantified by the noise standard deviation σ. Each noise level was run multiple times, except for single-run cases which were excluded. Table 1 summarizes the key metrics: average accuracy, signal efficiency, background efficiency, and final significance.

13.1 Clean Data ($\sigma = 0$)

The classifier demonstrates strong and stable performance on noise-free data. Accuracy ranged from 86.32% to 89.28%, with an average of 87.79%. Signal efficiency remained high (0.740–0.802), while background efficiency was very low (0.009–0.014). Final significance ranged from 6.79 to 7.49.

13.2 Low Noise ($\sigma = 0.05$)

Introducing small noise minimally impacted accuracy, which ranged from 86.92% to 89.80% (average 88.32%). Signal efficiency remained stable (0.751–0.819), and background efficiency stayed low (0.004–0.020). The final significance displayed considerable variance (5.85–12.20), suggesting that small noise can occasionally enhance separation in some runs but also introduces instability.

13.3 Moderate Noise ($\sigma = 0.1$)

At $\sigma = 0.1$, accuracy remained comparable to lower noise levels (86.28–89.92%, average 87.82%). Final significance showed a wider range (5.47–11.89), with occasional high values, but the trend begins to show degradation in performance stability. Signal and background efficiencies remain moderately consistent.

13.4 High Noise ($\sigma = 0.5$)

Performance begins to degrade at higher levels. Accuracy decreased to 83.40–87.44% (average 86.16%), while signal efficiency decreased and background efficiency increased, reflecting a reduction in classifier confidence. Final significance ranged from 3.96 to 5.24, indicating that high noise impairs the ability to reliably distinguish signal from background in gradient boosted algorithms.

13.5 Extreme Noise ($\sigma = 1.5$)

Classifier performance collapses under extreme noise. Accuracy dropped to 73.64–75.96% (average 75.28%), and signal efficiency fell below 0.66 on average. Background efficiency increased substantially, and final significance approached 1.69–1.80, which is consistent with near-random output classifications.

13.6 Overall Trend

The results indicate that the classifier is robust against low levels of noise ($\sigma \leq 0.1$), with minimal impact on accuracy and separation metrics. Intermediate noise ($\sigma \approx 0.05$) introduces variability in final significance, highlighting a regime where luck based generation of the gaussian put on top of the data can occasionally improve separation. Moderate and high noise levels ($\sigma \geq 0.5$) lead to consistent degradation in performance, with extreme noise ($\sigma = 1.5$) rendering the classifier almost wholly ineffective due to the extreme levels of accuracy drops (ΔAvg Accuracy$_{\text{Noise=0}}$ − Avg Accuracy$_{\text{Noise=1.5}} \rightarrow 12.51$) and final significance (ΔFinal Significance$_{\text{Noise=0}}$−Final Significance$_{\text{Noise=1.5}} \rightarrow$ 5.43) compared to lower levels of noise.

14 Discussion

Evidently, gradient-boosted algorithms such as XGBoost are robust to low levels of Gaussian noise, as they maintain high accuracy and high confidence in signal-background separation when $\sigma \leq 0.1$. This is interesting as at these noise levels, fluctuations in the data have minimal effect on classifier performance, as seen in Table 1 which indicates stable signal efficiency and low background efficiency. This was an unexpected result in the hypothesis as the expectation was that as the detector noise increases the accuracy would decrease as well. This is true to some extent but not to the extent that it acts linearly in that as noise decreases as does the accuracy...however this can be explained as the algorithm attempting to seek out relative similarities in train-test datasets. Therefore it would be unsurprising that at a point where noise reaches even higher that the algorithm finds it almost impossible to truly decipher signal and background events.

Intermediate noise ($\sigma \approx 0.05$) introduces variability in final significance across runs. While average accuracy remains high, some runs show enhanced separation, suggesting that small noise can occasionally vary accuracy due to the amount of values being separated randomly. This allows the classifier to better distinguish overlapping signal and background distributions. However, this

effect is unstable and highly dependent on the specific noise realization that comes from random distribution. So, although the accuracy may be higher, the variability in sequential runs is also higher indicating less consistency. This is consistent with the explanation for the dynamic changes in accuracy as $Noise_{\sigma}$ increases.

As noise increases further ($\sigma \geq 0.5$), the performance of the classifier is significantly reduced. Accuracy and signal efficiency heavily decreases while background efficiency rises, reflecting the reduced confidence of the classifier (XGBoost) in distinguishing events. At extreme noise levels ($\sigma = 1.5$), the classifier approaches random guessing, with a final significance near 1. This confirms the idea that detector resolution can critically limits machine learning performance in high-energy physics context.

In general, these findings illustrate the trade-off between algorithm durability and sensitivity in ML-based classification of particle collisions. They highlight the importance of realistic noise modeling when developing and testing classifiers for experimental datasets.

15 Conclusion

This research explores a core challenge in analyzing LHC datasets: distinguishing rare signal events from abundant background fluctuations under imperfect detector conditions. Gradient-boosted algorithms such as XGBoost exhibit strong performance under low noise but degrade as Gaussian smearing increases. Accuracy and ROC AUC remain nearly constant at low noise, then decrease progressively as noise increases, with a sharp decline at high noise levels. This demonstrates the sensitivity of classification performance to detector resolution. These findings highlight where classification remains reliable and when noise overwhelms the algorithms ability to separate background and signal effectively. This was indicated in the results to be around when $\sigma > 0.1$. This study provides a simplified but at the very least informative demonstration of how noise impacts ML-based classification in an high energy physics hadron collider context for gradient boosted algorithms. It brings attention to the importance of detector resolution when

collecting data for physicists and illustrates the possible implication that data, if not scrutinized, could spread misinformation. This research holds as a precedent for different algorithms separate from XGBoost or other gradient boosted algorithms and it is possible that there may be other algorithms such as a neural network or deep learning that may be even better for classification under gaussian smear. It also indicates the importance for physicists to take care and develop good noise models when developing or testing classifiers as being able to detect more signal collisions through high level classifiers would decrease the time it takes to find extremely important collisions, ultimately, increasing efficiency and saving even more time.

16 References

1. ATLAS Collaboration. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC.

2. Bartlett, D. (n.d.). CERN. https://indico.cern.ch/event/318531/attachments/612850/843143/daniela_l5.pdf

3. Genovese, D., et al. (2025). Mixture-of-expert graph transformers for interpretable particle collision detection. *Scientific Reports, 15*(1).

4. Giasemis, F. I. (2025). Real-Time Analysis of Unstructured Data with Machine Learning on Heterogeneous Architectures. *arXiv:2508.07423*.

5. Guardian. (2025). AI to revolutionize fundamental physics.