Indian Journal of Pure & Applied Physics
Vol. 63, December 2025, pp. 1147-1159
DOI: 10.56042/ijpap.v63i12.22747

A Comparative Study of BDT and DNN Algorithms for Higgs Boson Prediction

Manil Khatiwada[a*], Nabin Bhusal[a], Manjeet Kunwar[a], Krishna Baduwal[a] & Rajendra Neupane[b]

[a]Central Department of Physics, Tribhuvan University, Kathmandu 446 18, Nepal

[b]Department of Physics, Birendra Multiple Campus, Tribhuvan University, Bharatpur 442 00, Nepal

Received:14th August 2025; accepted:3rd November2025

The Higgs boson, also known as the "God Particle," is responsible for giving mass to elementary particles. Detecting and studying its production remains a major challenge in particle physics. In this study, we use deep neural networks and decision-boosted trees to identify the decay of the Higgs boson into four leptons. Our dataset includes millions of simulated collision events from the Large Hadron Collider. Results show that decision-boosted tree models are highly effective in recognizing complex patterns, improving the accuracy of Higgs boson detection. We evaluate model performance using key metrics such as ROC-AUC curves, Cross-Validation, background-to-noise ratio, and score distribution analysis. Our findings offer a strong framework for advancing Higgs boson research in high-energy physics.

Keywords: Higgs boson, Boosted decision tree (BDT), Deep neural network (DNN), Large hadron collider (LHC)

1 Introduction

The Higgs boson, first proposed by Peter Higgs in the 1960s, plays a crucial role in explaining how elementary particles acquire mass. Its discovery was a major milestone in particle physics, providing essential support for the Standard Model (SM). The ATLAS[1] and CMS[2] collaborations at the Large Hadron Collider (LHC) experimentally confirmed the existence of a particle with a mass around 125 GeV, consistent with the predicted Higgs boson. The LHC primarily produces Higgs bosons through gluon-gluon fusion, vector boson fusion, Higgs-strahlung with W or Z bosons, and associated production with top quarks (ggF, VBF, VH, and ttH, respectively)[3].

Gluon-gluon fusion (ggF) is the dominant Higgs boson production mechanism at the Large Hadron Collider (LHC), contributing significantly to the observed Higgs events. This process, where two gluons fuse to create a Higgs boson, is crucial for Higgs detection. The discovery of the Higgs boson relied heavily on observing its decay products in events produced via ggF[4,1]. While the Higgs boson decays through various channels, those with Lepontic final states, such as $H \rightarrow ZZ \rightarrow 4l$ (where l represents a lepton, e or μ), offer particularly clean signatures for detection due to the relatively low background[1,5-6]. Although the branching ratio for these Lepontic decays is smaller compared to other channels like $H \rightarrow \gamma\gamma$, the clear signals with well-measured leptons make them essential for precise measurements of Higgs properties and for searches for new physics[7].

Despite its importance, detecting the Higgs boson is extremely challenging due to its rarity, occurring in only one in a billion LHC collisions, and its short lifespan[8]. Moreover, the presence of overwhelming background noise makes it difficult to distinguish Higgs events from other Standard Model processes. Achieving a statistically significant observation required surpassing the 5σ threshold, which was made possible by precise classifiers only[9].

Traditional rule-based methods for Higgs boson detection faced severe limitations due to the vast volume and complexity of data. These approaches lacked the flexibility to capture non-linear relationships and intricate event features, making them impractical for effective classification. As a result, advanced computational techniques, particularly machine learning (ML), emerged as powerful alternatives. ML algorithms can efficiently process large datasets, uncover hidden patterns, and improve signal-to-noise discrimination, making them invaluable in modern particle physics research[10].

The significance of machine learning in Higgs boson classification was demonstrated in the 2014 Higgs Boson Machine Learning Challenge, where 12 ML-based algorithms significantly improved event classification accuracy[11]. Among these, Boosted Decision Trees (BDTs) and Deep Neural Networks

*Corresponding author: E-mail: khatiwadamanil111@gmail.com

(DNNs) have been widely explored. BDTs excel in handling structured data and are effective in various Higgs search strategies, while DNNs offer superior performance by automatically extracting complex features from raw data. Previous work[12] showed that DNNs can outperform BDTs in event classification, and more recent studies have demonstrated their success in distinguishing between vector boson fusion and gluon–gluon fusion Higgs production.

We focus on the Lepontic decay channel $H \rightarrow ZZ^* \rightarrow 4l$, where the Higgs boson decays into two Z bosons, which subsequently decay into four leptons (four electrons, four muons, or two electrons and two muons). This channel, with a branching ratio of 2.64×10^{-2}, provides a clean experimental signature. However, it is also affected by significant background contributions, including lepton pairs originating from photon conversions, Z + jets events with misidentified jets, and top-antitop tt processes where leptons emerge from semi-Lepontic heavy-flavour decays. These backgrounds necessitate advanced classification techniques for efficient event selection and analysis.

To address these challenges, we conduct a comparative study of BDT and DNN classifiers for Higgs boson event classification. BDTs, such as XGBoost[13] are effective for structured data and feature engineering, while DNNs offer a flexible framework for modeling complex event characteristics. Our study evaluates these methods based on classification performance, robustness to background noise, and overall predictive power.

The rest of this paper is organized as follows: Section 2 describes the data distribution, proposed algorithms, and evaluation methods. Sections 3 and 4 present the results and discussion, while Section 5 concludes with an overview of future work and challenges.

2 Material and Methods

2.1 Distribution of Higgs Dataset

The dataset employed in this study comprises 27 Higgs features as independent variables and one dependent response variable categorized into two classes: signal and background events. These features represent various kinematic and angular properties of leptons and jets, including lep_pt_0, lep_pt_1, lep_pt_2, lep_pt_3, deltaR_l0_l1, deltaR_l0_l2, deltaR_l0_l3, deltaR_l1_l2, deltaR_l1_l3, deltaR_l2_l3, and invariant_mass_4lep, among others. For the present analysis, however, only three key parameters: lep_pt_0, deltaR_l0_l1, and invariant_mass_4lep were selected to explore the properties of the Higgs boson. The selection of these features was guided by both their physical significance and statistical discriminative power.

Specifically, lep_pt_0 (the leading lepton transverse momentum) captures the energy scale of leptons produced in Higgs decays; deltaR_l0_l1 (the angular separation between the two leading leptons) reflects the event topology, with Higgs events exhibiting more collimated lepton pairs than background processes; and invariant_mass_4lep (the invariant mass of the four-lepton system) serves as the primary observable corresponding to the Higgs boson mass peak around 125 GeV. A correlation and feature-importance analysis revealed that many of the remaining variables were either redundant or contributed minimally to the model's predictive performance. Therefore, focusing on these three representative parameters ensured a balanced approach, enhancing model interpretability, avoiding overfitting, and preserving the essential physical characteristics necessary for accurate Higgs signal identification.

2.2 Kendall Heatmap Plots

The Kendall correlation heat map is showing the relationships between various input variables related to Higgs boson decay shown in Fig. 1.

The Kendall correlation matrix reveals strong positive correlations between the transverse momenta (p_t) of the four leptons (lep_pt_0, lep_pt_1, lep_pt_2, lep_pt_3), indicating a tendency for high p_t in one lepton to be accompanied by high p_t in the others. This is further reflected in the moderate positive correlation observed between the lepton p_ts and the invariant mass of the four-lepton system (invariant_mass_4lep), which is expected given that the invariant mass is derived from the lepton momenta. Conversely, the ΔR variables ΔR_{ij}, representing the angular separation between leptons, exhibit weak or near-zero correlations with the individual lepton p_ts. However, within the ΔR variables themselves, strong negative correlations are present, suggesting a relationship where a large separation between one pair of leptons is likely associated with smaller separations between other pairs, possibly constrained by geometry or the underlying particle decay production dynamics.

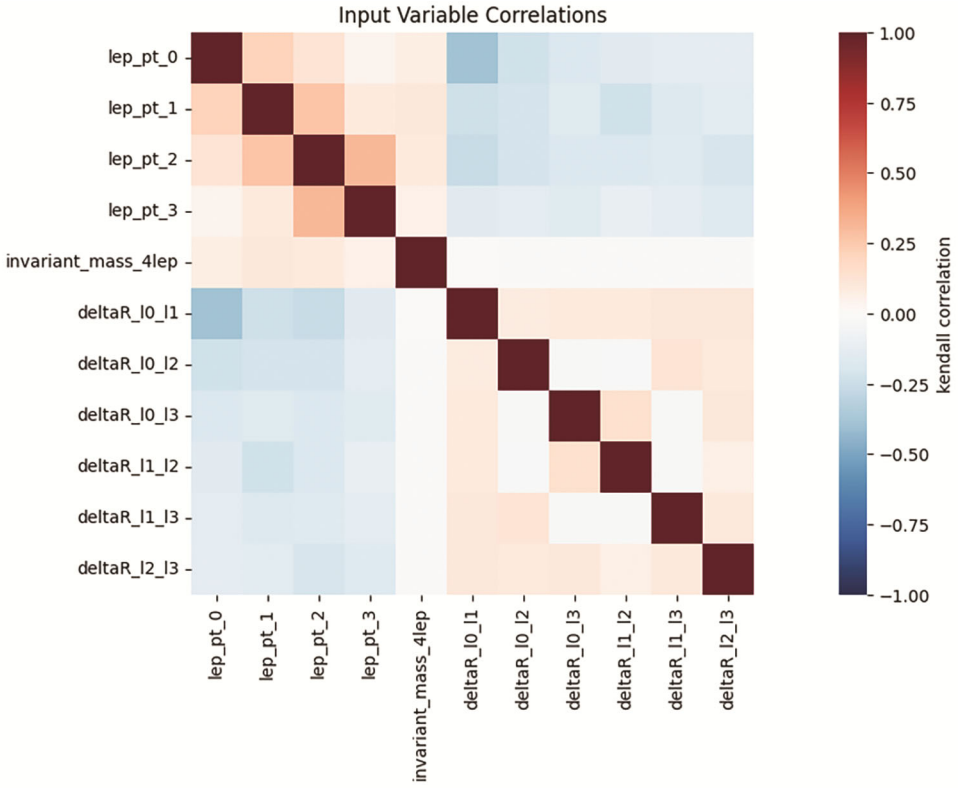

Fig. 1 — Correlation matrix of input variables, with red for negative and blue for positive correlations

Whereas the Kendall correlation heat map played an important role in selecting the final parameters for analysis. It helped identify which variables carried unique information and which were redundant due to high mutual correlations. Based on this, lep_pt_0, deltaR_l0_l1, and invariant_mass_4lep were chosen because they represent distinct and complementary physical aspects: momentum, angular separation, and invariant mass, while remaining largely independent of each other. This ensured that the selected features contributed meaningful, non-overlapping information to the model and improved both its interpretability and performance.

2.3 Deep Neural Network (DNN)

A deep neural network (DNN) comprises multiple interconnected layers, where each layer performs a linear transformation followed by a nonlinear activation function. Mathematically, the operation of the l^{th} layer can be expressed as

$$a^{(l)} = f\left(W^l a^{(l-1)} + b^l\right) \quad \ldots (1)$$

where $a^{(l)}$denotes the activation (output) of layer l, $W^{(l)}$and $b^{(l)}$represent the weight matrix and bias vector of the layer, respectively, and $f(\cdot)$is the activation function. Thus, the parameters $a$and $b$correspond to the layer activations and bias terms within the network.

$$Z^{(l)} = W^l a^{(l-1)} + b^l,\ a^l = f(Z^l) \quad \ldots(2)$$

During forward propagation, each layer transforms its input through this mapping. The activation function $f(z)$is chosen as the Rectified Linear Unit (ReLU) for hidden layers and the sigmoid function for the output layer. The network is trained using the binary cross-entropy loss function, where class weights $W$are incorporated to address data imbalance[14].

Optimization is performed using the Adam algorithm, which updates weights based on first and second moment estimates of the gradients. The updates are computed as follows:

$$m_l = \beta_l m_{(l-1)} + (1-\beta_l) g_t, v_t = \beta_2 v_{l-1} + (1-$$

$$\beta_2) g_t^2 W^{(l)} \leftarrow W^l - \eta \frac{\frac{M_t}{1-\beta_1^l}}{\sqrt{\frac{v_t}{(1-\beta_2^l)}} + \exists} \quad \ldots(3)$$

Where g_t represents the gradient at time step t.

During backpropagation, the gradients of the loss function with respect to the weights and biases are computed using the chain rule. The gradient of the loss with respect to the output layer is:

$$\frac{\partial \eta}{\partial Z^{(l)}} = \hat{y} - y \quad \ldots(4)$$

where $\hat{y}$ is the predicted output and y is the actual label. For hidden layers, the gradients propagate using:

$$\delta^{(l)} = (W^{(l+1)})^T \delta^{(l+1)} \circ f`\left(Z^{(l)}\right) \quad \ldots(5)$$

where $\delta^{(l)}$ represents the error term for layer l, and $f`\left(Z^{(l)}\right)$ is the derivative of the activation function. *The weight and bias updates follow*[15]:

$$W^{(l)} \leftarrow W^{(l)} - \eta \frac{\partial L}{\partial W^{(l)}}, b^{(l)} \leftarrow b^{(l)} - \eta \frac{\partial L}{\partial b^{(l)}} \quad \ldots(6)$$

where η is the learning rate.

The training of a neural network relies on backpropagation, a method used to adjust the model's parameters by propagating the error from the output layer back to the input layer. This process begins by computing the difference between the actual target values and the predicted outputs during the forward pass. The resulting error is then used to calculate the error gradient, which helps in updating the model's weights and biases to minimize prediction errors.

The optimizer plays a crucial role in this process, as it determines how the model updates its parameters. Popular optimization techniques include gradient descent, Adam (Adaptive Moment Estimation). These methods ensure that the model converges towards an optimal solution by iteratively refining the weights.

2.4 Boosted Decision Tree (BDT)

XGBoost follows a gradient boosting framework, where the final model F(x) is an ensemble of decision trees:

$$F(x) = \sum_{m=1}^{M} a_m h_m(x) \quad \ldots(7)$$

The model is trained iteratively, updating the prediction at each step:

$$F_m(x) = F_{m-1}(x) + a_m h_m(x)$$

The objective function consists of a loss term $l(y_i, f_i(x))$ and a regularization term $\Omega(h_m)$ to control complexity[16]:

$$\text{Obj}(\Theta) = \sum_i l(y_i, f_i(x)) + \sum_m \Omega(h_m) \quad \ldots(8)$$

For binary classification, the logistic loss function is commonly used[17]:

$$l\left(y_i, f_i(x)\right) = -[y_i \log\left(\sigma\left(F(x_i)\right)\right) + (1 - y_i)\log(1 - \sigma(F(x_i)))] \quad \ldots(9)$$

where $\sigma\left(F(x_i)\right) = \frac{1}{1+ e^{-F(x)}}$ is the sigmoid function.

The training of a Boosted Decision Tree (BDT) relies on an ensemble learning technique where multiple weak decision trees are sequentially trained, with each new tree attempting to correct the errors made by the previous ones. This process begins with a simple decision tree trained on the dataset, and the prediction errors are identified by assigning higher weights to misclassified samples. Subsequent trees focus more on these difficult cases, gradually improving the overall model accuracy. The final prediction is obtained by combining the outputs of all trees, often using weighted averaging or majority voting. Common boosting algorithms include Gradient Boosting, AdaBoost (Adaptive Boosting), and XGBoost (Extreme Gradient Boosting), each optimizing the model by minimizing a loss function. This iterative refinement helps Boosted Decision Trees achieve high predictive performance while reducing overfitting through techniques like learning rate adjustment and regularization.

2.5 Performance Evaluation Metrics

The ROC-AUC plot is used to measure the discriminative power of the classifiers. The Receiver Operating Characteristic (ROC) curve is generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at varying classification thresholds. These are defined as:

$$TPR = \frac{TP}{TP+FN} \text{ and } FPR = \frac{FP}{FP+TN} \quad \ldots(10)$$

where TP and TN represent the number of correctly classified positive and negative instances, respectively, while FP and FN denote the misclassified cases. The Area Under the Curve (AUC) quantifies the overall performance of the classifier and is computed as the integral of the ROC curve:

$$AUC = \int_1^0 TPR(FPR)d(FPR) \quad \ldots(11)$$

A higher AUC value indicates better separability between the two classes, with an ideal classifier achieving an AUC of 1.0, while a random classifier yields an AUC of 0.5.

In addition to the ROC-AUC metric, accuracy is computed to determine the overall correctness of the model's predictions. It is mathematically expressed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots(12)$$

Moreover, precision {also known as Positive Predictive Value (PPV)}, is an essential metric for evaluating the reliability of positive predictions. It is given by:

$$Precession = \frac{TP}{TP + FP} \quad \ldots(13)$$

A higher precision value indicates fewer false positives, making this metric particularly useful in applications where minimizing incorrect positive classifications is crucial. These performance metrics, along with the ROC-AUC plot, provide a comprehensive evaluation of the classification models, ensuring a robust assessment of their effectiveness in distinguishing between different classes. Looking to get an unbiased estimation of the performances of the algorithms, Higgs dataset from ATLAS is divided into a training sample (80 %) and test sample (20 %) in Table 1

3 Results and Discussion

3.1 Model Performance and Convergence Analysis

The network architecture consisted of three dense (fully connected) layers. The input layer comprised 64 neurons with ReLU activation, receiving input features of dimension n (where n represents the number of features in the training data). The subsequent hidden layer contained 32 neurons, also with ReLU activation. The output layer consisted of a single neuron with a sigmoid activation function, appropriate for binary classification. The model was compiled using the Adam optimizer and the binary cross-entropy loss function.

The binary cross-entropy (BCE) loss function quantifies the discrepancy between the predicted probabilities and the true binary class labels. It is mathematically defined as

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad \ldots(14)$$

where $y_i \in \{0,1\}$ denotes the true class label, and $\hat{y}_i$ represents the predicted probability that the instance belongs to class 1. Minimizing this loss function encourages the model to produce predictions that align closely with the true class labels, thereby improving classification performance during training.

Table 1 — Higgs datasets used to train the model

Event Parameter	Value
Total Signal Event	185020
Total Background Events	421740
Class (Signal, Background)	(s,b)

Performance was evaluated using accuracy and precision metrics. Training was conducted for 20 epochs with a batch size of 32, employing 20 % of the training data as a validation set. To address class imbalance, class weights were incorporated during training, specifically (0: 0.7699, 1: 1.4262). In the context of Higgs boson decay classification, class imbalance arises because the number of events corresponding to the signal (Higgs boson decays, class 1) is significantly smaller than the number of background events (class 0). This reflects the rarity of Higgs boson events in experimental data compared to background processes, which dominate the dataset. Applying class weights gives higher importance to the minority (signal) class, ensuring that the model does not become biased toward predicting the majority (background) class. Before training, the input features underwent standardization using the Standard Scaler. After training, the model achieved an accuracy of 0.9749 and a precision of 0.9528.

The 64–32–1 DNN architecture was selected after testing multiple configurations with 1–4 hidden layers and neuron counts ranging from 16 to 128. This design provided the best trade-off between complexity and accuracy, achieving high training accuracy with low validation loss. Larger networks exhibited overfitting, while smaller ones failed to capture the data's complexity. Additionally, Fig. 2 shows that 64–32–1 structure demonstrated consistent and stable performance across different random dataset splits.

In the given implementation, the XG Boost model is trained using the default hyperparameters. The maximum tree depth max_depth is set to 6, ensuring each tree does not become excessively complex. The learning rate learning rate is set to 0.3, controlling the step size for weight updates to prevent overfitting. The number of trees in the ensemble estimators is set to 100, allowing the model to learn iteratively while balancing performance and computational efficiency.

The maximum tree depth was chosen after evaluating values between 3 and 10. Trees with depths less than 5 were insufficient to capture the complex non-linear patterns in the data, whereas trees deeper than 7 tended to over fit without improving cross-validation performance. A depth of 6 offered the optimal balance between model complexity and generalization, delivering consistent results across multiple train/test splits.

The feature importance analysis from the BDT model highlights that invariant_mass_4lep is the most

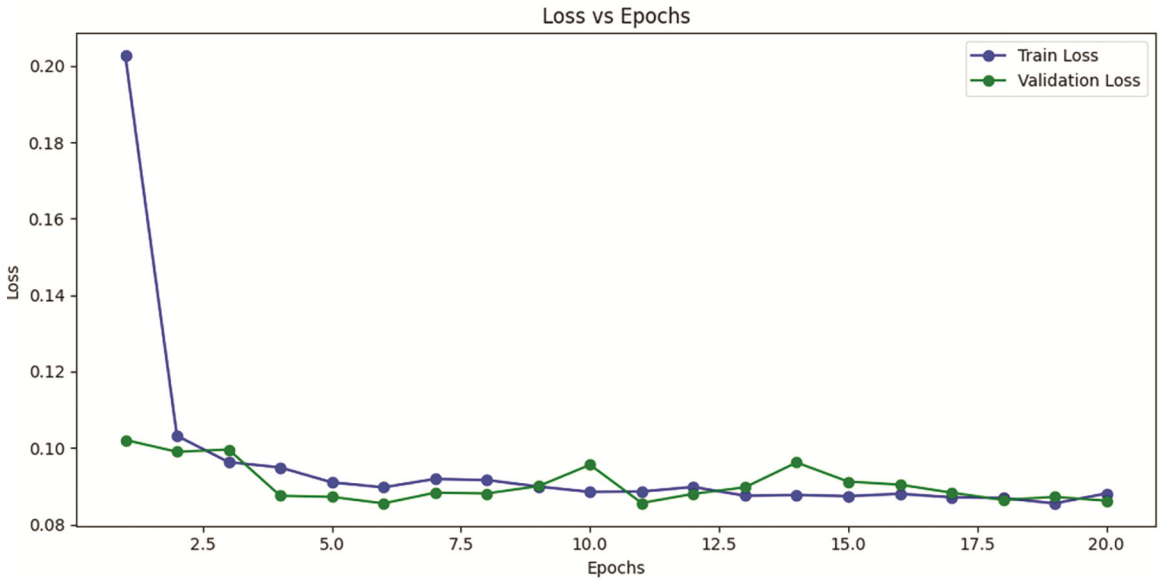

Fig. 2 — DNN convergence plot showing training and validation accuracy over 20 epochs, indicating stable learning

crucial variable, with a significantly higher importance score compared to all other features. This means that the total invariant mass of the four leptons has the strongest influence on the model's predictions. In contrast, other features such as transverse momentum p_t of individual leptons and angular separation ΔR between lepton pairs have much lower importance scores (Fig. 3). While these kinematic variables still contribute to the classification, their impact is minimal compared to the invariant mass. This suggests that the model relies heavily on the invariant mass to distinguish between different event types, making it the most valuable feature in the analysis[18]. Where a clearer separation between signal and background indicates a strong distinction between their distributions. However, in practice, several challenges complicate the calculation of invariant mass. One key issue is lepton misidentification, where hadrons or photons are wrongly reconstructed as leptons, contaminating the signal region[19]. Additionally, reconstruction inefficiencies, often arising from detector acceptance limits, resolution effects, or imperfect calibration, can bias the measurement. Incorrect lepton pairing further broadens and distorts the mass spectrum, particularly in multi-lepton final states where multiple possible combinations exist. Furthermore, final-state radiation (FSR) can cause leptons to lose part of their energy to emitted photons, shifting the reconstructed mass to lower values. Similarly, pile-up interactions in high-luminosity environments introduce extra tracks and calorimeter deposits, increasing the likelihood of false

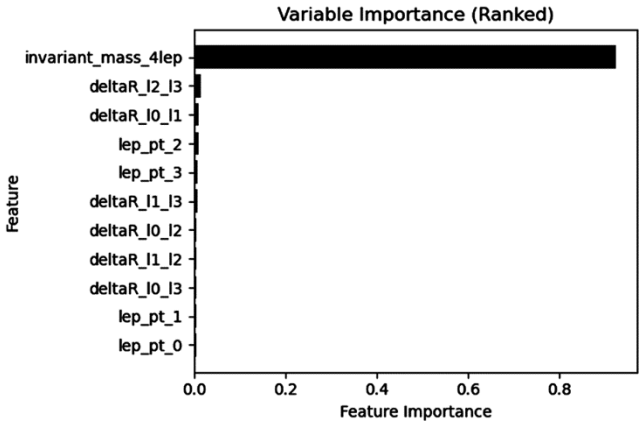

Fig. 3 — BDT variable ranking, showing the relative importance of features in the model's predictions

combinations. Collectively, these effects reduce the sharpness of the mass peak and make it harder to isolate the signal from background[20].

3.2 ROC AUC Curve Analysis

The ROC AUC analysis demonstrates excellent classifier performance for both the DNN and BDT models. The overall test AUC values are 0.934211 and 0.942311 for the DNN and BDT, respectively, indicating a strong ability to separate signal from background, which is crucial for Higgs boson detection. To evaluate generalization and overfitting, we plot the ROC curves for both the training and testing datasets (Fig. 4). A model that generalizes well will have training and test curves that are closely aligned. The BDT model shows near-perfect alignment between its training and test ROC curves. The DNN model, while still excellent, shows a

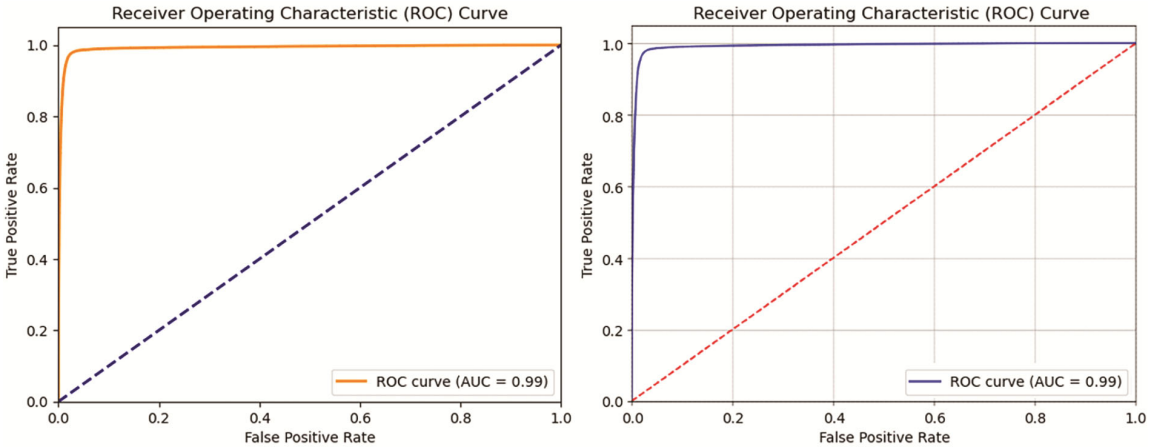

Fig. 4 — ROC curves for DNN (left) and BDT (right). DNN shows minor overfitting, while BDT aligns closely with better generalization and slightly higher test AUC. Diagonal dashed line indicates a random classifier

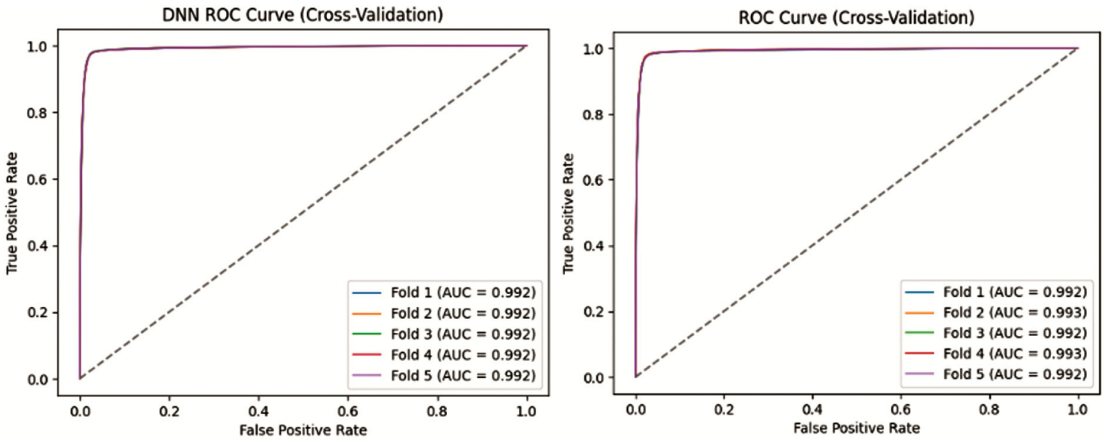

Fig. 5 — ROC curves from 5-fold cross-validation for DNN (left) and XGBoost (right) using uniformly processed background, signal, and real data

slightly larger gap between its training and test performance.

A higher AUC value suggests that the model effectively maximizes TP while minimizing FP and FN, thereby improving classification accuracy. The slight edge of the BDT over DNN implies that the BDT model showed better generalization in distinguishing between signal and background events. This visual evidence confirms that the BDT model generalizes slightly better, as it maintains its performance on unseen data more consistently than the DNN. This is further supported by its marginally higher test AUC. This careful monitoring of overfitting is necessary for validating discoveries like the Higgs boson in particle physics[21].

3.3 5-Fold Cross-Validation Results for the Two Classifiers on Datasets from CMS and ATLAS

Both the XGBoost and Deep Neural Network (DNN) classifiers were evaluated using 5-fold cross-validation on datasets obtained from the CMS and ATLAS experiments, shown in Fig. 5. To ensure uniformity and fairness in performance assessment, the background, signal, and real datasets were taken from the same distribution and processed under identical conditions

The XGBoost classifier was evaluated using 5-fold cross-validation to ensure robustness and generalizability, as shown in Table 2 in detail. The model achieved consistently high performance across all folds, with accuracy ranging from 0.9741 to 0.9764 and an average accuracy of 0.9752 ± 0.0009. Precision remained stable (mean = 0.9541 ± 0.0014), indicating the model's ability to correctly identify positive instances while minimizing false positives. The AUC values exceeded 0.992 in all folds, reflecting excellent separability between classes. The ROC curves from 5-fold cross-validation reveal that true negatives and true positives were significantly

Table 2 — 5-fold cross-validation results for XGBoost on CMS and ATLAS datasets using uniform background, signal, and real data

Fold	Accuracy	Precision	AUC	TN	FP	FN	TP
1	0.9751	0.9540	0.9921	66695	1740	889	36115
2	0.9764	0.9565	0.9930	66789	1646	843	36161
3	0.9748	0.9534	0.9924	66671	1764	888	36116
4	0.9755	0.9542	0.9931	66698	1737	850	36154
5	0.9741	0.9523	0.9923	66625	1809	918	36086

Table 3 — DNN performance across 5-fold cross-validation on CMS and ATLAS datasets, using consistently sourced and uniformly processed background, signal, and real data

Fold	Accuracy	Precession	AUC	TN	FP	FN	TP
1	0.9752	0.9589	0.9921	133791	3079	2144	71864
2	0.9750	0.9600	0.9920	133881	2989	2284	71724
3	0.9761	0.9590	0.9922	133792	3078	1962	72046
4	0.9755	0.9569	0.9922	133623	3247	1913	72095
5	0.9752	0.9594	0.9923	133832	3037	2186	71822

higher than misclassified instances, with false positives and false negatives remaining relatively low. These results demonstrate the model's strong predictive capability and reliability for the classification task, with minimal performance variance across folds.

The Deep Neural Network (DNN) model was assessed using 5-fold cross-validation to evaluate its generalization capability. Across all folds, the model-maintained accuracy between 0.9750 and 0.9761 as shown in detail in Table 3, with an average accuracy of 0.9754 ± 0.0004, indicating high stability in performance. Precision values ranged from 0.9569 to 0.9600 (mean = 0.9589 ± 0.0010), reflecting the model's effectiveness in correctly identifying positive samples while minimizing false positives. The AUC values were consistently above 0.992, with a mean of 0.9922 ± 0.0001, demonstrating excellent class separability. Analysis of ROC curves from 5-fold cross-validation showed that both true positive and true negative counts were substantially higher than misclassification counts, with relatively balanced false positive and false negative rates across folds. These results confirm the DNN's strong classification performance and robustness, with minimal variance in predictive capability across different data splits.

3.4 Score Distribution for Train and Test Samples

The comparative analysis of the Deep Neural Network (DNN) and Boosted Decision Tree (BDT) prediction distributions for signal and background events reveals distinct characteristics inherent to each model's architecture and learning process. BDT, shown in Fig. 6, exhibits a striking tendency towards extreme probability assignments, with predictions clustering near 0 and 1. This behaviour underscores BDT's strength in creating sharp, definitive decision boundaries, a hallmark of tree-based ensemble methods. This characteristic is well-documented in the literature, where BDTs, particularly gradient boosting variants like XGBoost, are recognized for their ability to efficiently partition feature space to minimize classification error[22].

The high concentration of predictions at the extremes suggests a model that is highly confident in its classifications, potentially reflecting its ability to isolate pure subsets of data within the feature space. However, this decisiveness comes with a potential caveat: the lack of predictions in the mid-range indicates a limited representation of uncertainty. Previous work[23] highlighted that gradient boosting, the algorithmic foundation of BDTs, focuses on iteratively correcting errors, resulting in models that prioritize strong classification signals over distinct probability estimates. This can lead to an overemphasis on binary decision-making, potentially overlooking subtle nuances or ambiguities within the data. Such sharp boundaries, while effective for clear separation, may also indicate a risk of overfitting, particularly if the training data is not fully representative of the broader population.

In contrast to the BDT, the DNN shows a distinctly different morphology in its prediction distributions, as illustrated in Fig. 6. Visually, the DNN histogram is characterized by a smoother and more continuous distribution of output scores, with a more pronounced population of events in the intermediate probability range (between 0.2 and 0.8). This observable difference in shape reflects the different learning architectures of the two models. BDT, an ensemble of decision trees, often produces distributions with

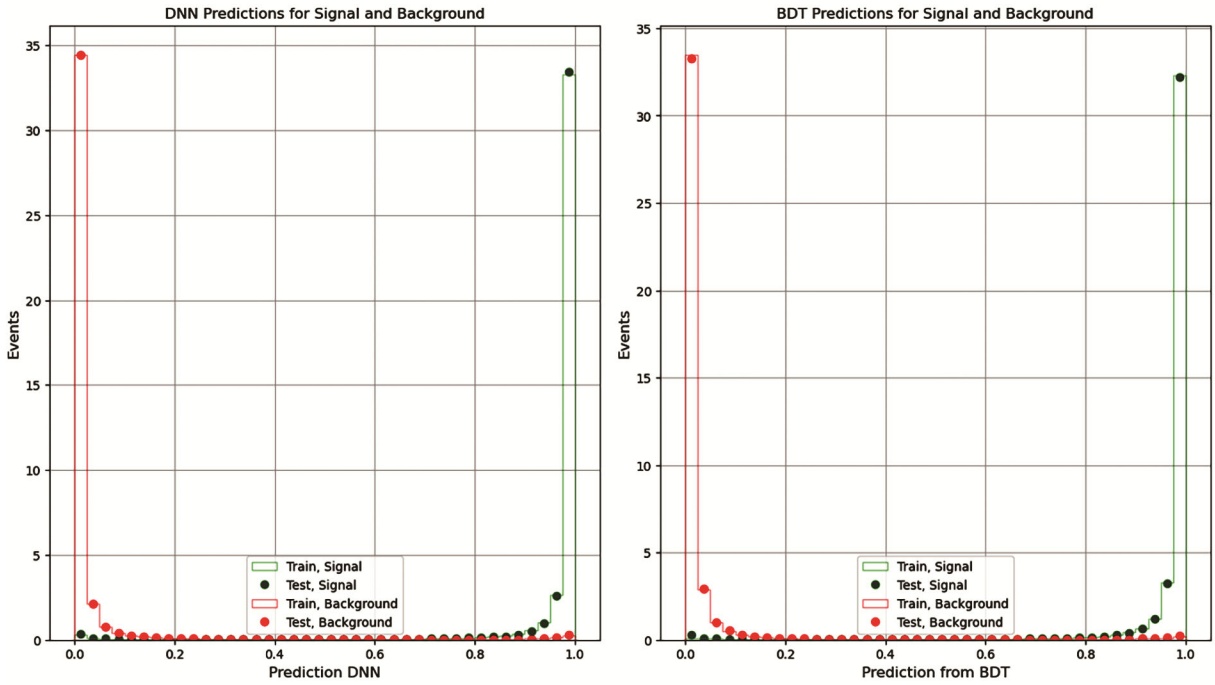

Fig. 6 — DNN (left) shows a smoother, broader score distribution, while BDT (right) has sharper peaks near 0 and 1. Both separate signal and background, but with different classification behaviors

sharper peaks at the extremes (near 0 and 1), which is consistent with its piecewise, binary splitting nature. The DNN, however, with its ability to model complex, non-linear relationships through multiple layers[24], generates a broader and more graded output spectrum. The presence of a substantial number of predictions in the intermediate range suggests that the DNN is capturing a more graded confidence in its classifications. This capability can be advantageous in scenarios where the distinction between signal and background is not always clear-cut, allowing the model to express varying degrees of certainty. Furthermore, previous work[25] has highlighted the ability of DNNs to quantify uncertainty through techniques such as Bayesian neural networks and dropout, which can provide valuable insights into model confidence and reliability. While the DNN also exhibits peaks near 0 and 1, indicating strong classifications, its broader distribution suggests a model that is potentially more robust and generalizable. However, this complexity and capability to capture uncertainty necessitate careful tuning to prevent overfitting, a common challenge in deep learning models. Ultimately, both models demonstrate effective separation of signal and background, indicating their utility in the given classification task. The choice between BDT and DNN hinges on the specific requirements of the application: the BDT for scenarios demanding clear, decisive classifications and the DNN for tasks where capturing uncertainty and distinct probabilities are paramount.

Table 4 — Evaluation of unseen datasets, ensuring uniformity by using background, signal, and real data

Quantity	Value
Total Signal Events	370040
Total Background Events	684349
Total Data Events	507

3.5 Model Evaluation and Validation for ATLAS and CMS Data

The efficacy of the trained Boosted Decision Tree (BDT) and Deep Neural Network (DNN) models in distinguishing signal from background events was evaluated using unseen datasets, as detailed in Table 4. The BDT, an ensemble learning method that combines multiple decision trees to improve prediction accuracy, and the DNN, a powerful model capable of learning complex data representations through hierarchical layers, were chosen for their proven ability to handle high-dimensional data and complex relationships inherent in particle physics analyses[24, 25].

The score distributions for signal, background, and data events were visualized using stacked histograms,

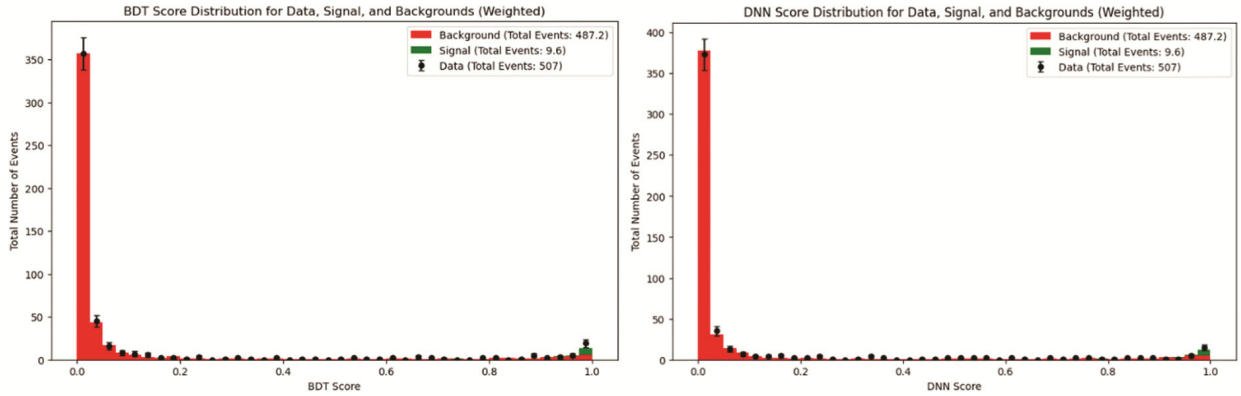

Fig. 7 — BDT (left) and DNN (right) score distributions for ATLAS data, showing weighted event counts.

as shown in Fig. 7. The BDT score distribution revealed a clear separation between signal and background, with the background concentrated at low scores and the signal peaking near 1. This demonstrates the BDT's ability to effectively discriminate between event types. The data distribution follows the background at low scores, but exhibits a shoulder at higher scores, suggesting the presence of signal-like events. Similarly, the DNN score distribution showed a distinct separation, with the background at low scores and the signal at high scores, highlighting the DNN's robust classification performance.

Both models were trained on weighted datasets to account for detection efficiencies, selection biases, and systematic uncertainties, ensuring an accurate representation of expected event rates[26]. The total number of events (signal: 9.6, background: 487.2, data: 507) underscores the rarity of signal events compared to background. The sharp peak of the signal distribution at high scores for both models indicates their strong selection power, crucial for enhancing signal significance in the presence of overwhelming background.

Table shows higher signal significance in the last two bins for BDT than DNN. To quantify the signal significance, we analyzed the last two bins of the histograms, where signal events are expected to dominate. The significance was calculated using the formula:

$$Z = \frac{S}{\sqrt{B}} \qquad \ldots(15)$$

where S is the signal count, and B is the background count in the relevant bins. The observed signal significance was then computed by replacing S with the excess of data over the background. For the BDT, the signal count in the last two bins was 8.408, the background count was 6.066, and the data count was 24.000. For the DNN, similar calculations were performed. The statistical uncertainty of the number of events in each bin was estimated using the Poisson distribution, with the standard deviation (error) given by $\sqrt{N}$, where N is the number of events in the bin.

Table 5 — Signal Significance Comparison: DNN vs BDT (Last Two ATLAS Bins)

Event Parameters	Deep Neural Network (DNN)	Boosted Decision Tree (BDT)
Signal Significant	1.381	2.033
Observed Significant	1.834	4.336

In the BDT output, a small fraction of background events appears at high scores, and some signal events extend into the mid-score range, indicating minor misclassification and overlap. In contrast, the DNN output shows a narrower background peak at low scores, a sharper clustering of signal events near 1.0, and minimal mid-score overlap, reflecting stronger discrimination power. This sharper separation in the DNN suggests higher background rejection efficiency and improved signal purity, which is advantageous for analyses requiring clean signal extraction. Additionally, the closer confinement of DNN scores toward the extremes (0 and 1) implies a higher confidence in classification decisions, potentially leading to better stability in downstream significance calculations. Overall, while both classifiers achieve effective separation, the DNN demonstrates a slight performance edge in purity, background suppression, and classification confidence shown in Fig. 8.

The successful application of BDTs and DNNs in this analysis aligns with their established use in high-energy physics, where they have played pivotal roles in discoveries such as the Higgs boson[1, 27-28]. These models effectively leverage complex feature spaces to enhance signal-to-background discrimination,

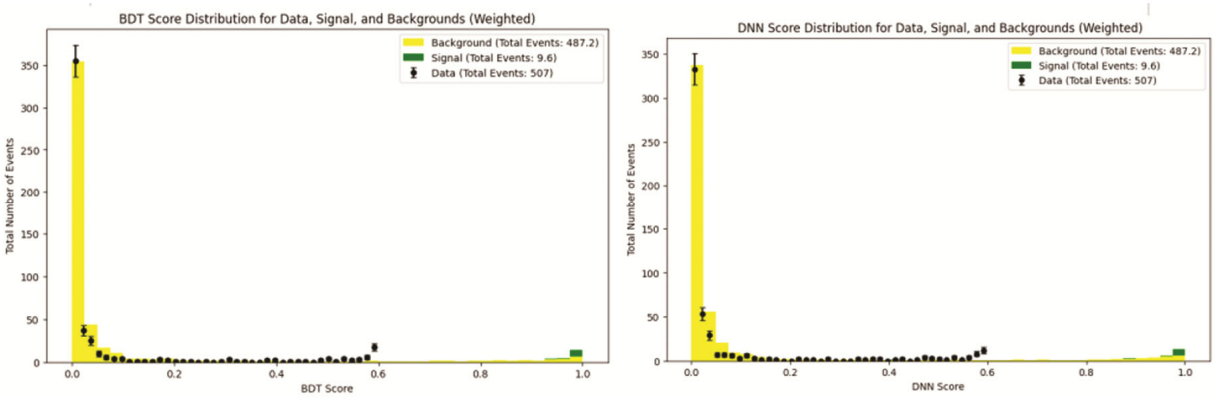

Fig. 8 — BDT (left) and DNN (right) score distributions for CMS data, showing weighted event counts. Last two bins have comparable signal significance, with DNN slightly higher

demonstrating their utility in extracting rare signals from noisy data.

3 Discussion

Comparative analyses between Boosted Decision Trees (BDTs) and Deep Neural Networks (DNNs) in Higgs boson decay event classification consistently highlight a trade-off between model confidence and generalization. BDTs, known for their ability to construct sharp decision boundaries, often exhibit higher confidence in classifications, manifesting as output scores concentrated near 0 and 1. This behaviour is attributed to the ability of BDTs to capture complex, non-linear relationships, particularly in kinematic variables critical for Higgs boson identification, such as transverse momentum and invariant mass, as confirmed by feature importance analyses. This aligns with established findings in high-energy physics, where BDTs have shown strong performance in feature-rich, complex datasets. Previous studies on Higgs boson discovery[27] demonstrated that BDTs offered robustness and interpretability, particularly in identifying signal-rich regions based on kinematic features, effectively exploiting the differences between signal and background events and showcasing their strength in high-energy physics classification tasks.

The sharp peaks observed in the BDT score distribution, while indicative of high confidence, raise concerns about potential overfitting, a known challenge with BDTs, particularly in high-dimensional datasets. In contrast, DNNs exhibit a smoother transition and a broader spread of prediction scores, suggesting improved generalization. This characteristic is advantageous for handling ambiguous events and enhancing robustness to unseen data. However, the overlap in DNN score distributions in the mid-range region highlights challenges in fully resolving subtle signal–background distinctions, necessitating further architectural optimization. These observations are consistent with previous studies[28], which emphasize the importance of deep architecture, regularization techniques, and careful hyperparameter tuning in optimizing DNN performance for high-energy physics applications, where complex data patterns require robust modelling.

Although Fig. 6 and Fig. 8 may visually appear to show only subtle differences between the DNN and BDT output distributions, a closer examination reveals distinct underlying behaviours that justify our claim of a more nuanced picture with the DNN. In Fig. 6, while both models achieve similar classification boundaries, the DNN's broader and smoother probability distribution indicates a more gradual transition between signal and background events, reflecting improved generalization and reduced overconfidence compared to the sharply peaked BDT outputs. This difference, though not striking in visual amplitude, becomes evident when analyzing the score density and spread, where DNN predictions exhibit a continuous gradient rather than abrupt separations. Similarly, in Fig. 7 and Fig. 8, the comparative score distributions show that BDT tends to cluster predictions toward extreme values (0 and 1), while DNN maintains a wider mid-range distribution, capturing ambiguous events more effectively. Thus, the subtle visual differences are statistically meaningful and consistent with the expected model behaviour, reinforcing that DNNs, despite producing smoother outputs, retain valuable discriminative power through improved handling of uncertain classifications.

Statistical hypothesis tests reveal a higher signal significance for BDT in the last two bins, underscoring its efficacy in identifying rare events. Specifically, the observed significance for ATLAS datasets indicates superior signal detection capability, as shown in Table 5. This consistency across unseen evaluation datasets reinforces the reliability of both models. The necessity for XGBoost implementation emphasizes the importance of handling large datasets for future ATLAS detector analyses. Additionally, exploring function classes that incorporate jet-related aspects presents a promising direction for enhancing signal-to-background discrimination. This approach aligns with ongoing efforts to integrate physical insights into machine learning models, improving both performance and interpretability. Previous work[29] has shown that incorporating jet substructure information can significantly enhance the sensitivity of Higgs boson analyses.

For the CMS data, the tests show that both the BDT and DNN models perform well in the last two bins, as shown in Table 6. However, DNN has a marginally better signal significance. This close performance suggests that both models are effective in detecting signals, each with its own strengths. Using methods like XGBoost can help handle large CMS datasets more efficiently. Also, adding features related to jets could help the models better distinguish signals from background noise. This fits with recent studies that show including physics-based information improves the results of Higgs boson detection.

Similarly, the analysis of signal significance in the last two bins, as presented in Tables 5 & 6 reveals a notable disparity between the Deep Neural Network (DNN) and Boosted Decision Tree (BDT) models. Specifically, the BDT demonstrates a significantly higher observed significance we compared to the DNN. This result aligns with previous studies in high-energy physics, where BDTs have demonstrated superior performance in identifying rare events and distinguishing signal from background. Notably, the initial Higgs boson discovery analyses reported[1,2] that BDTs effectively enhanced signal significance by exploiting complex kinematic correlations. The higher significance achieved by BDTs in this analysis underscores their capability to provide more robust signal detection, particularly in regions of phase space where signal-to-background discrimination is challenging, a characteristic often attributed to their sharper decision boundaries and effective feature exploitation.

Table 6 — Signal Significance Comparison: DNN vs BDT (Last Two CMS Bins)

Event Parameters	Deep Neural Network (DNN)	Boosted Decision Tree (BDT)
Signal Significant	2.021	1.929
Observed Significant	3.407	4.021

Similarly, comparing the results from ATLAS and CMS datasets reveals some notable differences in model performance. For ATLAS data, the Boosted Decision Tree (BDT) clearly outperforms the Deep Neural Network (DNN) with a higher observed significance of 4.336 versus 1.834, showing a stronger ability to detect rare signals. In contrast, the CMS data shows a more balanced performance between the two models: while BDT has a slightly higher observed significance (4.021 compared to 3.407 for DNN), the DNN achieves a marginally better signal significance (2.021 over 1.929 for BDT). This suggests that for CMS, both models have comparable strengths in identifying events, whereas for ATLAS, BDT holds a more distinct advantage. These differences highlight the importance of tailoring machine learning methods to the specific characteristics of each dataset and encourage further exploration of hybrid approaches and feature enhancements, such as incorporating jet-related variables to maximize signal detection across experiments.

4 Conclusion

This study conducted a comparative analysis of Boosted Decision Tree (BDT) and Deep Neural Network (DNN) classifiers for Higgs boson identification, utilizing datasets from both the ATLAS and CMS experiments. The objective was to evaluate which model and which dataset provides a clearer signal for Higgs boson detection.

Overall, the regularized BDT classifier demonstrated superior performance, achieving a higher AUC-ROC of 0.942311 compared to DNN's 0.934211. DNN's performance was notably hampered by a significant overlap between signal and background events in the score range of 0.4 to 0.6, indicating a region of ambiguity where its discriminatory power was less effective. Statistical uncertainties in critical physical variables, such as transverse momentum, energy, and invariant mass, further contributed to prediction variance, particularly in regions with low signal-to-background ratios.

A key finding from the cross-experiment comparison is that the performance and clarity of

Higgs boson detection are highly dependent on the specific dataset. For the ATLAS data, the BDT model was unequivocally superior, achieving an observed significance of 4.336, which was more than double that of the DNN (1.834). This suggests that the ATLAS data, when processed with a BDT, offer a much clearer and more significant signature of the Higgs boson. In contrast, for the CMS data, the models performed more comparably. The BDT still held a slight edge in observed significance (4.021 vs. 3.407), but the DNN achieved a marginally higher signal significance (2.021 vs. 1.929), indicating a more balanced trade-off between the models and a different underlying data structure.

References

1 Aad G, *et al*, *Phys Lett B*, 716 (2012) 1.
2 Chatrchyan S, *et al*, *Phys Lett B*, 716 (2012) 30.
3 Cho A, *Sci*, 6114 (2012) 1558.
4 Khachatryan V, *et al*, *Eur Phys J C*, 74 (2014) 3076.
5 Khachatryan V, *et al*, *Phys Lett B*, 765 (2017) 194.
6 Aaboud A, *et al*, *Eur Phys J C*, 76 (2016) 6.
7 Khachatryan V, *et al*, *Phys Rev Lett*, 114 (2015) 191801.
8 Chatrchyan S, *et al*, *Phys Lett B*, 713 (2012) 2.
9 Adam-Bourdarios C, Cowan G, Germain C, Guyon I, Kégl B & Rousseau D, in *NIPS 2014 Workshop High-Energy Phys & Mach Learn*, (2015) p. 19.
10 Baldi P, Sadowski P & Whiteson D, *Phys Rev Lett*, 114 (2015) 111801.
11 Carneiro T, Da Nóbrega R V M, Nepomuceno T, Bian G -B, De Albuquerque V H C & Reboucas Filho P P, *IEEE Access*, 6 (2018) 61677.
12 LeCun Y, Bengio Y & Hinton G, *Nature*, 521 (2015) 436.
13 Volkova S, in *3rd Int Conf Opt, Comput Appl & Mater Sci (CMSD-III 2023)*, 100 (2024).
14 Chen T & Guestrin C, in *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*, 785 (2016).
15 Azmi SS & Baliga S, *Int Res J Eng Technol*, 7 (2020) 6867.
16 Chatrchyan S *et al*., *Phys Rev D,* 89 (2014) 092007.
17 Han T, Parke S, *Phys Rev Lett*, 71 (1993) 1494.
18 Dicus D A & Willenbrock S D, *Phys Rev D*, 37 (1988) 1801.
19 Wilczek F, *Mod Phys Lett A*, 21 (2006) 701.
20 Friedman J H, *Ann Stat*, 29 (2001) 1189.
21 Breiman L, *Mach Learn*, 45 (2001) 5.
22 Vogt M, *at-Automatisierungstechnik*, 66 (2018) 690.
23 Cowan G, Cranmer K, Gross E & Routenburg O, *Eur Phys J C*, 71 (2011) 1.
24 Arpaia P, Azzopardi G, Blanc F, Buffat X, Coyle L, Fol E, Giordano F, Giovannozzi M, Pieloni T, Prevete R & Redaelli S, *IEEE Instrum & Meas Mag*, 24 (2021) 47.
25 Pang LG, *Int J Mod Phys E*, 33 (2024) 2430009.
26 Radovic A, Williams M, Rousseau D, Dragone M, Donkers J, Baltz E, *et al*, *Nature*, 560 (2018) 41.
27 Guest D, Cranmer K, Whiteson D & Yanez J, *Annual Rev Nucl Particle Sci*, 68 (2018) 161.
28 Butter A, *et al*, *J High Energy Phys*, 10 (2016)1.
29 Chatrchyan S, *et al*, *Phys Lett B*, 716 (1) (2012) 30.