Performance Investigation of BDTs for ω-muon Identification in CBM Experiment at FAIR

Summer Project

by

Raktim Mukherjee

Supervised by

Prof. Subhasis Chattopadhyay

Experimental High Energy Physics & Applications Group

Variable Energy Cyclotron Centre, Kolkata

August, 2023

1 Introduction

In the pursuit of understanding the fundamental building blocks of the universe, the field of High-Energy Physics (HEP) continually seeks innovative methods to analyse complex particle interactions. The Compressed Baryonic Matter (CBM) experiment at the Facility for Antiproton and Ion Research (FAIR) is a remarkable endeavour, poised to study the behaviour of strongly interacting matter under low temperatures and very high chemical potential conditions. The CBM experiment is expected to have a very high interaction rate, up to 10MHz [1].

To achieve these ambitious objectives, the CBM experiment employs heavy-ion collisions, a powerful tool to create the elusive quark-gluon plasma. By colliding two atomic nuclei at high energies, researchers create conditions reminiscent of the early universe, where quarks and gluons are no longer confined within hadrons. One particularly illuminating probe in this endeavour is the detection of di-leptons, which emerge as decay products of particles containing charm quarks and low-mass vector mesons. In this study, the primary focus lies in detecting the di-leptonic decay of the ω-meson, a critical exploration within the CBM experiment's framework.

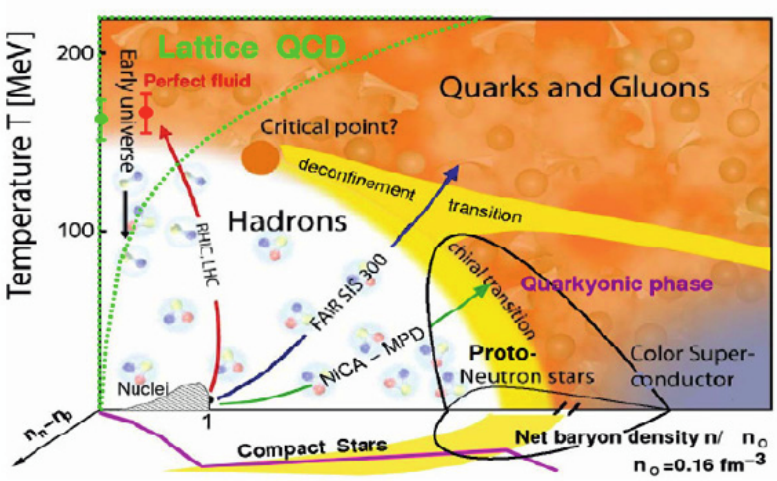

Figure 1: The phase diagram of strongly interacting QCD matter schematically showing the phase[2]

With its unique properties and characteristics, Muon identification has emerged as a crucial facet of particle physics research. The ω-meson, a light vector meson, plays a pivotal role in probing the behaviour of the strong force in heavy-ion collisions. Its interactions with other particles provide invaluable insights into the properties of the Quark-Gluon Plasma (QGP), a state of matter that existed microseconds after the Big Bang[3]. Accurate identification of ω-mesons and their decay products, including muons, amidst a complex background, is an intricate challenge that demands sophisticated analysis techniques.

In recent years, machine learning methods have proven their mettle in tackling intricate classification tasks, offering an alternative paradigm to traditional analysis techniques. Boosted Decision Trees (BDTs), an ensemble learning technique, have garnered significant attention for their ability to handle high-dimensional, non-linear data while providing interpretable results. This project aims to assess the performance of BDTs using multiple configurations in the context of ω-muon identification within the CBM experiment at FAIR.

1.1 Objective

The primary objective of this study is to comprehensively study the effectiveness of Boosted Decision Trees in identifying ω-muons amidst the intricate background noise inherent in CBM experimental data. This evaluation entails the examination of various aspects, including classification accuracy, efficiency and purity of signal, robustness against variations in the configuration of the trees, and the interpretability of the trained model. Two different boosting techniques have been used in this project - *Adaptive Boost* and *Gradient Boost*. The models have also been compared with our traditional analysis method- Rectangular Cuts.

2 Data Acquisition and Simulation

The dataset utilized in this investigation was generated through Monte-Carlo simulation, a crucial tool for approximating the behaviour of particle interactions within the experimental setup. For the CBM experiment, the signal dataset encompasses muons originating from the decay of the ω-meson, simulated using the PLUTO event generator(without embedding them in background). The background dataset involves muon-like tracks, simulated using the UrQMD (Ultra-relativistic Quantum Molecular Dynamics) model.

The track reconstruction process, which translates raw detector signals into meaningful particle trajectories, was accomplished using the CBM-ROOT framework. The following detector geometries were used
Silicon Tracking System (STS) version: *19A*
Muon Chamber version: *20A*

2.1 Feature Selection and Preprocessing

The model is trained using four features

1. `STS_hits`: The number of hits in the Silicon Tracking System
2. `MUCH_hits`: The number of hits in the Muon Chamber
3. `Chi_STS`: The standard deviation of the reconstructed path of the muon in the Silicon Tracking System
4. `Chi_MUCH`: The standard deviation of the reconstructed path of the muon in the Muon Chamber.
5. `Chi_VERTEX`: The standard deviation of the reconstructed path of the muon from the collision fireball.

Before delving into the training and evaluation of machine learning models, meticulous data preprocessing was conducted to ensure the quality and relevance of the dataset. Particles possessing `Chi_MUCH > 50` were excluded, as were particles that did not enter the Muon Chamber (`MUCH_hits = 0`).

The *transverse momentum*(p_T) and *rapidity*(y) were also obtained for the respective tracks but not used as features for training. This was done in order to divide our dataset into multiple p_T and y bins since the statistics, as well as the behaviour of the signal and background particles, vary at different ranges.

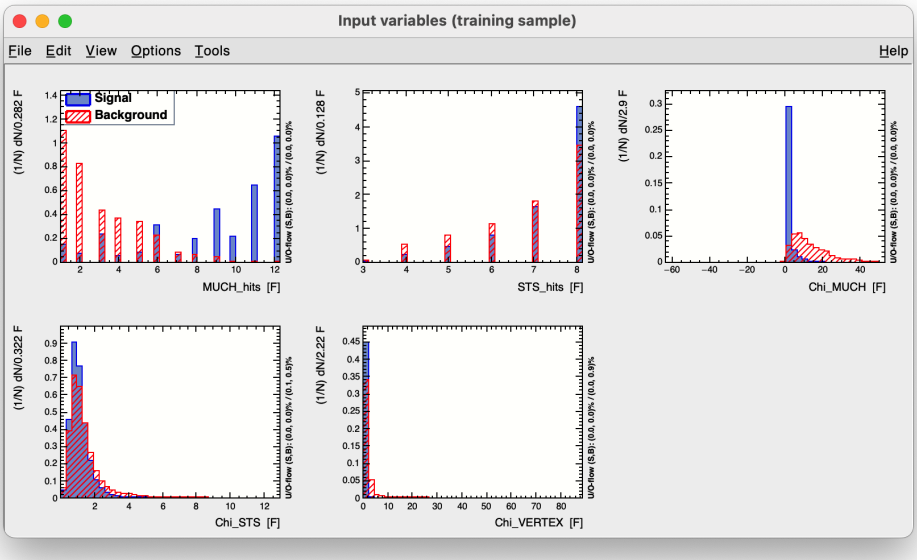

Figure 2: Preprocessed data used for training and testing

The refined dataset, representing the di-leptonic decay of the ω-meson, along with carefully selected features, sets the stage for the subsequent investigation into the performance of Boosted Decision Trees (BDTs) for ω-muon identification within the CBM experiment.

In the following section, we delve deeper into the intricacies of Boosted Decision Trees, their training process, and the methodology employed to evaluate their performance in identifying ω-muons amidst the complexities of experimental data.

3 Boosted Decision trees

Decision trees are versatile and intuitive machine-learning models used for both classification and regression tasks. These models represent a flowchart-like structure where each internal node represents a decision based on a particular feature, leading to one of its child nodes. At the leaf nodes, predictions are made based on the majority class (for classification) or a mean value (for regression) of the instances falling into that node.

Boosting, on the other hand, is an ensemble learning technique that combines the predictive power of multiple weak learners (usually simple models) to create a strong learner. Boosting iteratively trains a sequence of models, with each subsequent model focusing on correcting the mistakes made by the previous ones. This results in a final model with enhanced accuracy. The details of Adaptive boosting can be found in [4] and for Gradient boosting in [5].

The parameters varied for this project and how they affect the performance are as follows:

No. of Trees The number of individual decision trees in the ensemble. Increasing the number of trees can improve model performance, but it may also increase computation time and risk overfitting.

Min. Node Size The minimum number of samples required to create a new node in a decision tree. Larger values can simplify the tree and prevent overfitting, but overly large values may result in underfitting.

Beta The learning rate in gradient boosting. It controls the contribution of each tree to the final prediction. Smaller values make the model more robust but may require more trees for high accuracy.

Shrinkage Also known as the learning rate, shrinkage controls the step size in gradient boosting. Smaller values require more trees for high accuracy but often yield better generalization and robustness.

Bagged Sample Fraction The fraction of the training dataset randomly

selected for each tree in the ensemble. A lower fraction introduces randomness and can reduce overfitting.

nCuts The number of potential cut points considered when splitting a node in a decision tree. Increasing nCuts may lead to more complex trees and potential overfitting.

Max Depth The maximum depth or levels of a decision tree. A deeper tree can capture more complex relationships but may overfit. Controlling max depth helps prevent overfitting.

Separation Type The criterion used to evaluate the quality of a split when building a decision tree. Common types include Gini impurity and misclassification error. The choice can impact tree structure and, consequently, model performance.

3.1 Importance in HEP

Decision trees and boosting techniques have found extensive use in the field of High-Energy Physics [4] [6] due to their compatibility with the unique characteristics of particle physics data:

- **Complexity**: Particle physics data often exhibits complex relationships that decision trees can effectively capture, allowing for accurate classification and identification.

- **Interpretability**: Decision trees provide insights into feature importance, aiding physicists in understanding the underlying physical processes.

- **Handling Noise**: Decision trees can handle noisy data and are robust to irrelevant features, both of which are common in HEP experiments.

While simple cuts can offer quick and interpretable solutions, multivariate analysis techniques like BDTs are better suited for handling complex and high-dimensional data. They provide improved classification performance, increased robustness, and enhanced adaptability, making them a valuable tool for challenging classification tasks in HEP and other domains. Further details can be found in [7].

3.2 Application to CBM Experiment

In the context of the CBM experiment, adaptive boosting and gradient boosting algorithms have been employed for ω-muon identification. These BDT variants enhance the identification of ω-muons from complex backgrounds, contributing to the CBM experiment's goals. The remarkable performance of BDT for the same task in the GRAPES-3 experiment at TIFR [8] further justifies the choice of model for this project.

4 Methodology and Results

Approximately, 66000 background and 70000 signal tracks were used from the simulation. The total number of tracks was divided 50-50 into training and testing sets.

The Boosted Decision Trees with different configurations and boosting methods were created using TMVA [9], as shown in Tables 1 and 2. The models are also compared with traditional Rectangular Cuts. The Cuts as suggested by TMVA after cut-optimisation are

`MUCH_hits` > 4.12036

`STS_hits` > 3.74152

`Chi_MUCH` $<= 3.24484$

`Chi_STS` $<= 15.0062$

`Chi_VERTEX` $<= 14.087$

Corresponding to the efficiency of 80%. Obviously, the number of hits must be rounded off to an integer since the machine converts the data to float.

For the purpose of comparison, the performance of the models on the entire dataset was checked as can be seen from figure 3.

The decision trees were subsequently subjected to training and testing using data categorized into distinct p_T and y bins, as illustrated in Figure 4. Subsequently, each model, including the Rectangular Cuts approach, was applied to each specific data range, and their performance was compared to the results obtained using the entire dataset. The efficiency depicted in the graphs corresponds to the point where the significance, represented as $\frac{S}{\sqrt{S+B}}$,

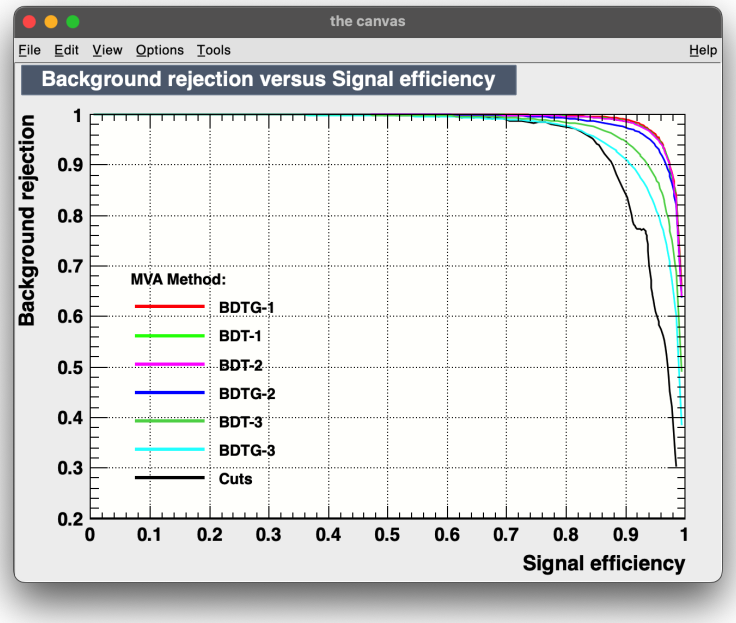

Figure 3: ROC curve of the different models. Area under the curve gives us the maximum purity obtainable from the model.

Parameters	**Config-1**	**Config-2**	**Config-3**
No. of Trees	1000	850	1200
Min. Node Size (%)	2.5	2.5	5
Beta	0.3	0.5	0.4
Bagged Sample Fraction	0.5	0.5	0.5
nCuts	20	20	25
Max Depth	3	3	3
Separation Type	Gini	Gini	Misclassification Error

Table 1: Parameter Configurations for Boosted Decision Trees (Adaptive Boost)

Parameter	**Config-1**	**Config-2**	**Config-3**
No. of Trees	1000	1000	1200
Min. Node Size (%)	2.5	2.5	5
Shrinkage	0.1	0.1	0.08
Bagged Sample Fraction	0.5	0.5	0.5
nCuts	20	20	20
Max Depth	5	2	3
Separation Type	Misclassification Error	Gini	Gini

Table 2: Parameter Configurations for Boosted Decision Trees (Gradient Boost)

attains its maximum value. Purity values indicate the highest achievable purity for each model, as measured by the area under the Receiver Operating Characteristic (ROC) curve. The final accuracy metric is provided by the TMVA framework.

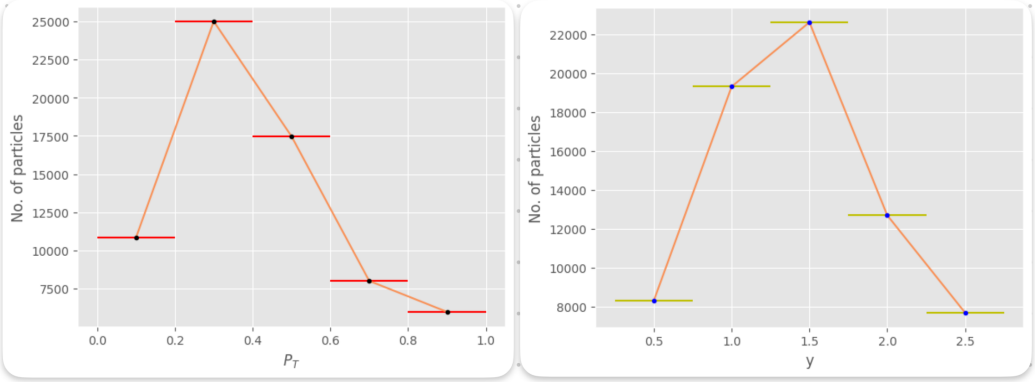

Figure 4: The bins considered for p_T and y respectively

From the data presented in Figure 5, a noticeable decline in performance becomes evident within the momentum range of 0.2 to 0.4 GeV/c. This observed deviation is unexpected, given that higher momentum ranges typically exhibit improved track separability. To investigate this anomaly further, the data within this specific range was subdivided into two adjacent bins, resulting in the first two bins being defined as $p_T < 0.3$ GeV/c and 0.3 GeV/c $\leq p_T < 0.6$ GeV/c. Surprisingly, this division still reveals a discernible dip within this region, as illustrated in Figure 6.

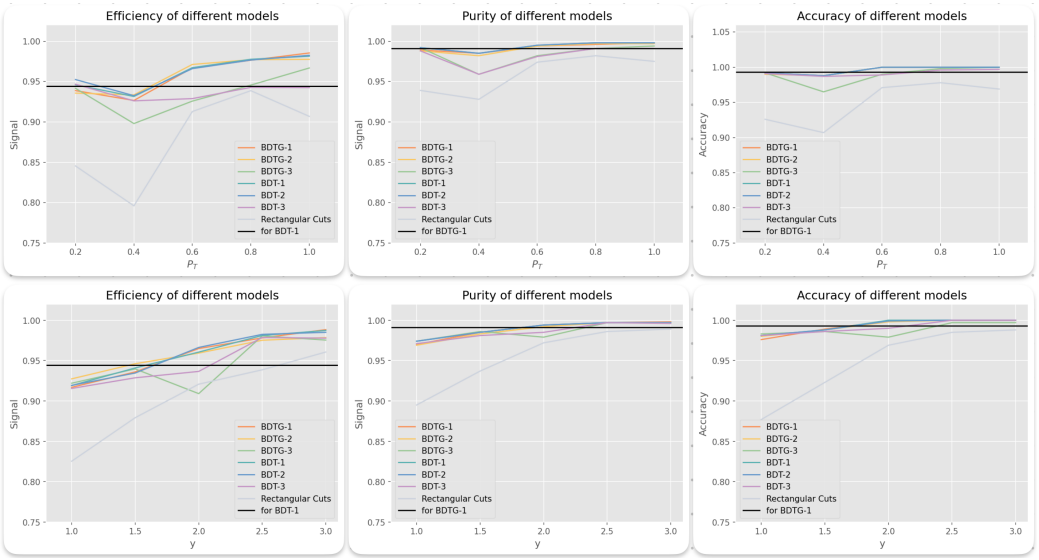

Figure 5: The performance of various models at different momentum (in GeV/c) and rapidity ranges. The black line represents the highest value obtained after training the models for the entire range

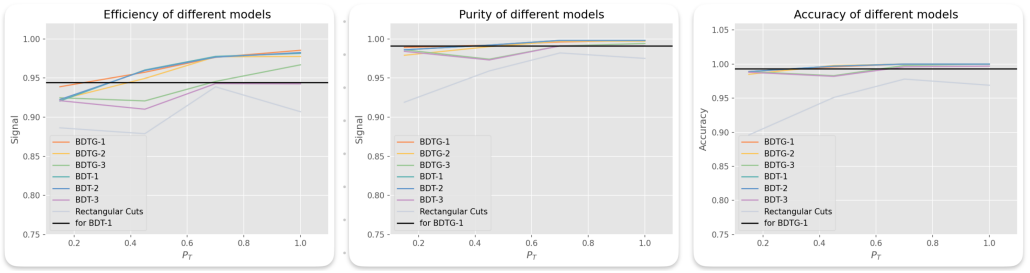

Figure 6: The performance of various models at different momentum (in GeV/c) ranges with different binning from Figure 5

Since the performance of the models is well above 90%, the models were trained over the entire range by excluding each variable once and using the other 4 for training and testing. From Figure 7 we can see that excluding `Chi_VERTEX` and `Chi_MUCH` results in a noticeable change in the performance of our models.

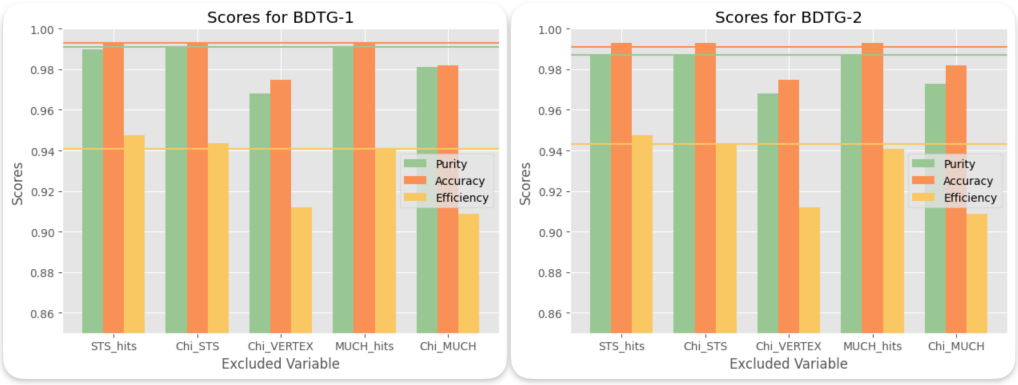

Figure 7: The performance of BDTG-1 and BDTG-2 the entire momentum range by excluding one variable. A similar trend was seen for the other models as well.

5 Conclusion and Discussions

Selecting the optimal model from our ensemble poses a challenge, given that, with the exception of BDT-3 and BDTG-3, discernible differences in performance among the other models are minimal, as evidenced in Figure 5. Such behaviour might be caused by the number of trees (both having 1200). Notably, our results show the substantial superiority of machine-learning algorithms over conventional Rectangular Cuts. However, for $y > 2.0$, we observe a convergence in purity and accuracy between the two techniques. This suggests that, under specific conditions, traditional methods can yield comparable results.

The performance dip observed within the momentum range of 0.2 to 0.4 GeV/c lacks a clear explanation. One plausible hypothesis is a substantial disparity between signal and background counts in this region.

Our models' commendable accuracy attests to their robustness against misclassification. Moreover, the minimal ($< 5\%$) divergence between testing and training errors serves as compelling evidence of their resistance to overfitting. Notably, as transverse momentum and rapidity values increase, our models approach near-perfect efficiency, purity, and accuracy. This phenomenon likely arises from the limited statistics available, leading to stringent classification criteria and consequently near-perfect outcomes.

These models offer the potential for constructing invariant mass spectra of the ω-meson. They can also be adapted to other decay channels, such as the di-electronic decay of the ω-meson or the identification of additional particles like J/ψ.

Furthermore, the models' performance can be further enhanced by incorporating embedded signals, a modification that would introduce a more realistic context into the analysis. This adjustment holds promise for refining the models' performance in real-world scenarios.

Acknowledgements

I am deeply appreciative of the support and guidance provided throughout this project. My sincere gratitude goes to my supervisor, **Prof. Subhasis Chattopadhyay**, who not only served as a source of inspiration but also provided invaluable guidance that brought out the best in me. Collaborating with him on this endeavour offered me valuable insights into the practices and methodologies of High-Energy Physics. I am truly thankful for his unwavering support, even amidst his busy schedule.

I extend my heartfelt thanks to **Dr. Nabhiraj PY** and **Dr. Ranjini Menon** for affording me the opportunity to work at VECC (Variable Energy Cyclotron Centre) and for their consistent encouragement throughout the duration of this project. My experience at VECC has been exceptionally rewarding.

During my time at VECC, I had the privilege of interacting with numerous outstanding individuals who have not only been colleagues but have also become friends. I would like to express my appreciation to **Mr. Abhishek Sharma** for generously providing the project data, without which my work would have been significantly hindered. **Dr. Biswarup Paul** played an instrumental role in helping me draw conclusions and provided invaluable suggestions for improvement. The unwavering support and encouragement from **Mr. Arun K Yadav** and **Dr. Partha Pratim Bhaduri** have been instrumental throughout my journey, and I am grateful for their willingness to assist whenever I sought their guidance.

References

[1] P. P. Bhaduri, "The physics goals of the CBM experiment at FAIR," *PoS*, vol. CPOD2021, p. 031, 2022.

[2] V. Kekelidze, A. Kovalenko, R. Lednicky, V. Matveev, I. Meshkov, A. Sorin, and G. Trubnikov, "Nica complex and jinr - status and plans," *EPJ Web of Conferences*, vol. 70, 03 2014.

[3] S. Chattopadhyay, Y. Viyogi, P. Bhaduri, and A. Dubey, "Participation in the compressed baryonic matter experiment at fair," 03 2011. [Online]. Available: https://www.currentscience.ac.in/Volumes/100/05/0682.pdf

[4] Coadou, Yann, "Boosted decision trees and applications," *EPJ Web of Conferences*, vol. 55, p. 02004, 2013. [Online]. Available: https://doi.org/10.1051/epjconf/20135502004

[5] J. Son, I. Jung, K. Park, and B. Han, "Tracking-by-segmentation with online gradient boosting decision tree," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[6] S. Khan, V. Klochkov, O. Lavoryk, O. Lubynets, A. I. Khan, A. Dubla, and I. Selyuzhenkov, "Machine learning application for {*upLambda* hyperon reconstruction in CBM at FAIR," *EPJ Web of Conferences*, vol. 259, p. 13008, 2022. [Online]. Available: https://doi.org/10.1051%2Fepjconf%2F202225913008

[7] C. Böser, S. Fink, and S. Röcker, "Introduction to boosted decision trees a multivariate approach to classification problems," Indico@KIT. [Online]. Available: https://indico.scc.kit.edu/event/48/contributions/3410/attachments/1690/2312/BDT_KSETA_Freudenstadt.pdf

[8] D. Bezboruah, M. Chakraborty, M. Devi, S. Dugad, S. Gupta, B. Hariharan, Y. Hayashi, J. Jagadeesan, A. Jain, P. Jain, S. Kawakami, H. Kojima, S. Mahapatra, P. Mohanty, Y. Muraki, P. Nayak, T. Nonaka, A. Oshima, D. Pattanaik, and M. Zuberi, "Machine learning model for separation of

muons from punch-through hadrons in eas at grapes-3 experiment," 07 2023, p. 522.

[9] A. Hocker *et al.*, "TMVA - Toolkit for Multivariate Data Analysis," 3 2007.