Contribution ID: **714** Contribution code: **contribution ID 714** Type: **Poster**

# Explainability of High Energy Physics events classification using SHAP

Understanding the predictions of a machine learning model can be as important as achieving high performance, especially in critical application domains such as health care, cybersecurity, or financial services, among others. In scientific domains, the model interpretation can enhance the model's performance, helping to trust them accurately for its use on real data and for knowledge discovery. Explainable artificial intelligence (XAI) is the current research field that proposes methods and techniques for producing more explainable models, and for understanding the predictions of machine learning models [1].

In the High Energy Physics (HEP) field, the complicated nature of the physics processes and data has required the use of complex machine learning models, like tree ensembles with thousands of trees, or deep neural networks with thousands of layers and millions of parameters [2]. These complex machine learning models are viewed as *black-box* systems that lack transparency and interpretability for getting a better model performance.

This work is focused on the use of the SHapley Additive exPlanations (SHAP) [3] —a post-hoc explainability technique of the XAI field— for interpreting machine learning classification models of High Energy Physics (HEP) data. The SHAP method is based on the coalitional game theory [4], where groups of players make decisions as coalitions building cooperative behavior. The goal is to distribute the total gain, or payoff, among players, according to the relative importance of their contributions to the final outcome of a game. Shapley values provide a solution to the assignment of a fair or reasonable reward to each player. In the machine learning context the *game* is the output of the model, and the *players* are the features included in the model.

SHAP assigns importance values (so-called SHAP values) that summarize the importance of each feature on the model prediction, contributing to the model transparency (as opposed to the *black-box* system). Briefly speaking, the SHAP value of each feature is computed as the average marginal contribution of a feature value across all possible coalitions [5].

In this work, we use the publicly available **Higgs dataset** [6]. This is simulated data, and the problem is to identify the signal from the background, where the signal corresponds to a Higgs boson decaying to a pair of bottom quarks according to the process: $gg \rightarrow H^0 \rightarrow W^{\mp}H^{\pm} \rightarrow W^{\mp}W^{\pm}h^0 \rightarrow W^{\mp}W^{\pm}b\bar{b}$.
Each event is represented by a set of 28 features, including 21 low-level features corresponding to physics properties measured by the detector, and 7 high-level features derived from the previous ones.

The **methodology** of this work includes the building of machine learning classifiers for the identification of signal and background (binary classification), using eXtreme Gradient Boosting (XGBoost) and deep neural networks (DNN). We select the classifiers with the highest performance for analyzing the local (and global) explainability using SHAP. The best XGBoost classifier achieved F1-score 0.74, precision 0.73, recall 0.76, accuracy 0.73, and AUC-ROC 0.72. The best configuration was obtained on the training process, including hyperparameter optimization, using cross-validation, 10-fold, and the XGBoost Python module. The best DNN classifier achieved F1-score 0.68, precision 0.67, recall 0.69, accuracy 0.66, and AUC-ROC 0.71. This network was built using mainly two Python modules: Keras for neural network model development and Talos [7] for automatic hyperparameter tuning, including the optimization of the DDN architecture parameters.

The **interpretability stage using SHAP approach** included the use of the Python SHAP module [8]. We calculated the SHAP values in each model, using the TreeExplainer and DeepExplainer methods for XGBoost and DNN, respectively. Considering SHAP is a post-hoc approach, both explainers have as input the previously build classification model and the testing dataset. The output of the method is the list of SHAP values, which quantify the feature impact on the classifier prediction of each event of the dataset. The best form to

see the SHAP values and how each feature impact on the event classification is visualizing plots, like the force plot, the dependence plot, and the summary plot, among others (current plots are available at [9]).

The force plot shows the contribution of each feature to the final score of the classifier. For instance, the XGBoost prediction of the first event (labeled as background, predicted as background) of the testing dataset is highly influenced by $m_{wbb}$, $\text{jet1}_{\text{eta}}$, $m_{wwbb}$ , and $m_{bb}$ features, where $m_{wbb}$ and $\text{jet1}_{\text{eta}}$ push the prediction lower, which makes sense, considering the output score of a XGBoost classifier is in the range $[0, 1]$, and scores close to 0 is expected to be in the background class.

Tree-based models provide global measures of feature importance (based on the total gain). However, computing the SHAP values across the whole dataset improves the gain-based measure, by taking the SHAP values over the whole dataset, and keeping the theoretical principles of SHAP values. The SHAP *summary plot* provides this information and shows a feature rank in descending order, and also a plot showing the magnitude and prevalence of a feature's impact. From our experiments, the top variables obtained for the XGBoost classifier were $m_{bb}$, $m_{wwbb}$, $m_{wbb}$, $m_{jjj}$, $m_{jlv}$, $jet1_{pt}$. For the DNN case, the top variables were $m_{bb}$, $m_{wwbb}$, $m_{wbb}$, $m_{jjj}$, $m_{jlv}$, $jet1_{pt}$. We can observe, that the high-level features belong to the top ranking, and hence, they are contributing more to the model prediction.

To the best of our knowledge, the use of the SHAP method in the context of High Energy Physics is recent, and only a few reports in the literature are applying this method in the HEP scenario. This is an ongoing work, and the future tasks include the development of a framework able to explain different machine learning models using SHAP, using datasets of other physics phenomena of interest. In addition, SHAP values can be used as a feature selection technique.

---


**References**

1 Alejandro Barredo Arrieta et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion, 58:82–115, 2020.

2 Dan Guest, Kyle Cranmer, Daniel Whiteson. Deep Learning and its Application to LHC Physics. Ann.Rev.Nucl.Part.Sci. 68 (2018), 161-181.

3 Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

4 L.S.S. Shapley, "A value for n-person games," in Contributions to the Theory of Games, 1953, pp. 307–317.

5 Molnar, Christoph. Interpretable machine learning. A Guide for Making Black Box Models Explainable, 2019.

6 https://www.openml.org/d/23512

7 https://github.com/autonomio/talos

8 https://github.com/slundberg/shap

9 https://github.com/rpezoa/hep_shap


## Significance

Nowadays, the understanding of a machine learning model can be as crucial as the prediction's accuracy of the model, especially in critical tasks such as disease diagnosis, automated vehicles, or financial services. In the High Energy Physics field the understanding of complex models allows us to trust them accurately to identify physics of interest, and draw conclusions against proposed theories. In addition, the interpretability of the model prediction could enhance the classifications of events, especially in the context of imbalanced classification, where the class of interest (the signal) is significantly underrepresented compared to the background.

## References

## Speaker time zone

Compatible with America

**Primary authors:** TORRES, Claudio Esteban (Federico Santa Maria Technical University (CL)); PEZOA RIVERA, Raquel (Federico Santa Maria Technical University (CL)); SALINAS, Luis (UTFSM)

**Presenter:** PEZOA RIVERA, Raquel (Federico Santa Maria Technical University (CL))

**Session Classification:** Posters: Raspberry

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools