The Federation – A novel machine learning technique applied to data from the Higgs Boson Machine Learning Challenge

Maximilian Mucha, Eckhard von Toerne
Physikalisches Institut, University of Bonn, Germany
E-mail: max.mucha@uni-bonn.de, evt@physik.uni-bonn.de

Abstract. The Federation is a new machine learning technique for handling large amounts of data in a typical high energy physics analysis. It utilizes UMAP to create an initial low-dimensional representation of a given set of trainings data. The dataset in this representation is clustered by using HDBSCAN. These clusters can then be used for a federated learning approach, in which we separately train one classifier for each cluster on the high-dimensional data. As a requirement for this approach, we need to apply an Imbalanced Learning method [9] to the data in the found clusters before the training. By using a Dynamic Classifier Selection method, the Federation can then make predictions for the whole dataset. As a proof of concept for this novel technique, open data from the Higgs Boson Machine Learning Challenge [1] is used and comparisons to results from established methods will be presented.

1. Introduction

The current research focus in high energy physics (HEP) is either refining measurements of the Standard Model or discovering evidence for physics beyond it. This is done by analyzing the massive amounts of data generated by the LHC each year and can be a challenging task even for modern machine learning (ML) algorithms. To tackle this issue, a novel machine learning technique called the Federation has been developed. The Federation allows for the training of machine learning models on large datasets without being constrained by computing limitations.

2. The Federation

Concept The Federation is a technique which uses a federated learning approach to tackle the issue of training machine learning models on very large datasets by dividing them into N subsets, each representing a distinct region of the feature phase space of the dataset. For each subset, one model is trained, resulting in an expert ensemble of N models, with each data subset and model, forming a Federation member (FM). The autonomy of the FMs allows for the learning process to be distributed to independent computing nodes, reducing the computational overhead of each model. The Federation consists of a ***common*** part and a ***diverse*** part, with the former analyzing the global structure of the feature phase space to create FMs, while the latter is formed by the FMs where each FM learns only a local part of the entire dataset. The name "Federation" was chosen to reflect the synergy between these two parts in achieving the goal of making predictions for large datasets. The choice of the ML model for the Federation members

is entirely up to the user, making this novel technique suitable not only for high energy physics but also for other fields where large amounts of data require processing via machine learning models. Therefore, this technique can be used to address a diverse range of ML problems, including binary classification, multi-class classification, and regression.

Training process The training process for the Federation, depicted in figure 1, starts with the ***common*** part, which consists of a combination of a non-linear dimensional reduction technique and a clustering algorithm. The primary objective of this part is to uncover hidden structures

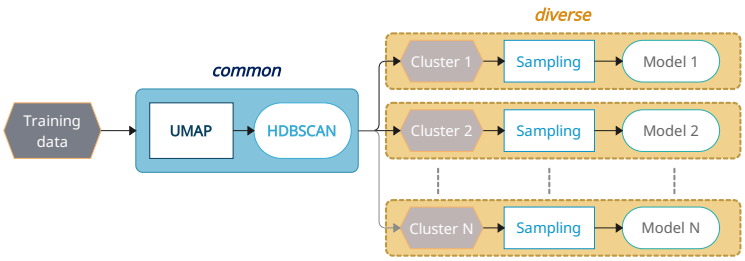


Figure 1: This figure depicts the training process of the Federation, where UMAP is utilized to transform the high-dimensional training data into a low-dimensional embedding. The resulting embedding is then clustered using HDBSCAN. UMAP and HDBSCAN form the ***common*** part of the Federation, which partitions the training data into N clusters. The ***diverse*** part is made up of N Federation members, each of which is composed of one portion of the clustered data and one ML model.

in the dataset and allocate a cluster label to each data point based on these patterns. Two commonly used non-linear dimensional reduction techniques are t-SNE [2] and UMAP [3]. These unsupervised machine learning methods aim to preserve the local and global structure of the high-dimensional data while projecting it into a low-dimensional representation. Through this process, data points with similar features are placed in close proximity to one another, thereby facilitating the formation of clusters in the resulting embedding. A demonstration of this concept for the MNIST dataset [4] is depicted in figure 2.

(a) (b)

Figure 2: Visualization of the MNIST handwritten digits dataset [4]: (a) Example digits with 784 pixels per images (b) Two-dimensional UMAP representation of the MNIST handwritten digits dataset. Each point represents the image of a handwritten digit. All points are colored according to the digit they represent.

In comparison to t-SNE, UMAP has the additional capability of predicting the location of new, unseen high-dimensional data points in a low-dimensional representation using a pre-

learned embedding. This feature is important for the Federation prediction, making UMAP the selected non-linear dimensional reduction technique. However, UMAP does not provide information about which points belong to which formed clusters. Therefore, a clustering algorithm must be performed on the low-dimensional UMAP embedding to obtain a cluster index for each data point. We have selected HDBSCAN [6], a density-based clustering algorithm because it is able to predict to which cluster a new unseen data point should be assigned based on a pre-learned clustering. The resulting number of clusters is dynamically determined by HDBSCAN.

The ***diverse*** part of the Federation is constructed by segmenting the input data based on the identified clusters, resulting in N distinct FMs. Every member of the Federation consists of the input data associated with its respective cluster and its model instance. Due to the possibility of class label imbalance resulting from the clustering process, the Federation offers the option to apply Imbalanced Learning techniques [9] to the FM training data to mitigate the decline in predictive accuracy.

Federated predictions UMAP and HDBSCAN can be used in combination as a Dynamic Classifier Selection (DCS) [12] method. By using UMAP to transform a new data point into the pre-learned embedding, and subsequently applying HDBSCAN to determine the most appropriate cluster for the new data point based on its position in the low-dimensional representation a cluster index can be assigned to a new unseen data point. The Federation can then select the FM that is the expert of the feature space of this new data point and utilize the corresponding ML model to make predictions, as illustrated in figure 3. By utilizing this DCS method for each data point, the Federation can make predictions in a federated manner for an entire dataset.

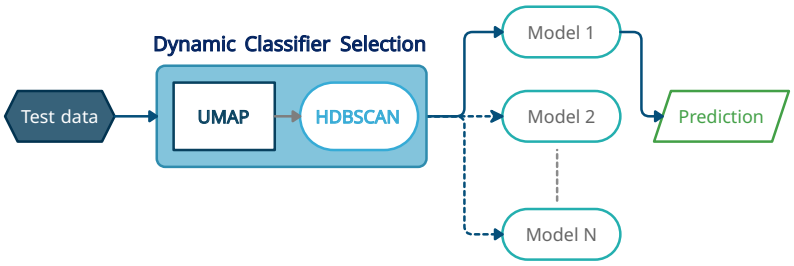


Figure 3: Representation of the DCS process for choosing the expert model of an Federation member for predicting test data. Initially, the high-dimensional test data is transformed into a low-dimensional embedding via a pre-learned UMAP. The resulting low-dimensional embedding is then clustered by a pre-learned HDBSCAN model. The cluster indices predicted by HDBSCAN are subsequently used to dynamically select the appropriate ML model from the Federation member, which acts as an expert for the given test data point.

Hyperparameters There are several UMAP variants available, we opt for the supervised parametric version [5], as it yields the best outcomes as part of the DCS method [14]. The UMAP embeddings are influenced by two hyperparameters (`n_neighbors` and `m_dist`) that determine the balance between the preservation of local versus global structure and the degree of tightness of data points in the lower dimensional embedding. Our research showed that these two parameters can have a significant impact on the prediction performance of the Federation [14]. The final significant hyperparameter that affects the UMAP embedding is the low-dimensional embedding's dimension. For visualization purposes, the dimension was set to two. To determine

the hyperparameters for HDBSCAN, Bayesian optimization [7] is performed during the training process where the DBCV index [8] is utilized as a scoring mechanism. Although no Imbalanced Learning techniques (over/undersampling) are applied by default, it is possible to incorporate sampling techniques [10] during the training process.

3. Application of the Federation

We have chosen the Kaggle Higgs Boson Machine Learning (HiggsML) Challenge [1] as an example application to evaluate if the idea of the Federation is feasible and can be used for analysis in HEP. This challenge provides publicly available ATLAS data consisting of simulated proton-proton collisions. Each single collision event is marked as signal $\mathcal{S}$ ($H \rightarrow \tau\tau$ decay) or background $\mathcal{B}$ (other decays) event. Thus, the challenge formulates a binary classification problem. Each event is represented by a data point with 30 features (17 kinematic and 13 derived). An additional challenge compared to other commonly used datasets in ML is that this dataset includes a categorical feature (the number of detected jets per event) as well as features which can have undefined values. Each data point comes with an event weight[1], which is used as sample weight. The event weights correspond to data taken in 2012. The sum of the signal weights $\sum_{i \in \mathcal{S}} w_i = N_s$ and the sum of background weights $\sum_{i \in \mathcal{B}} w_i = N_b$ are fixed constants which correspond to the expected total numbers of signal and background events, respectively. Therefore, whenever a subset of the training or the validation data is constructed, the event weights of the subset have to be renormalized according to these constants. The training dataset was used for training and the public and private leaderboard datasets were used for validation and testing, respectively.

To compare the predictive performance of the Federation with the HiggML challenge winners the Approximate Median Significance (AMS) was chosen as a figure of merit. The goal of the challenge was to maximize the AMS score. No additional features engineering was applied during our research.

3.1. Baseline analysis

During the HiggsML challenge the use of XGBoost was very popular [1]. For comparability reasons, we have therefore chosen XGBoost as the baseline model. The hyperparameters used in our study are based on [13]. During training, the event weights were used as sample weights. Because of the class imbalance of the data[2] the AUC of the Precision-Recall curve (AUC-PR) was used as an evaluation metric during training [11]. This singular XGBoost classifier is our single classifier baseline. An additional baseline was defined to compare the Federation clustering approach with a similar but simpler approach. This second cluster-based baseline clusters the datasets according to the categorical jet feature of the data points. An expert model was then trained for each cluster. Based on the value of the categorical feature of an unseen data point, an expert model is then selected for prediction.

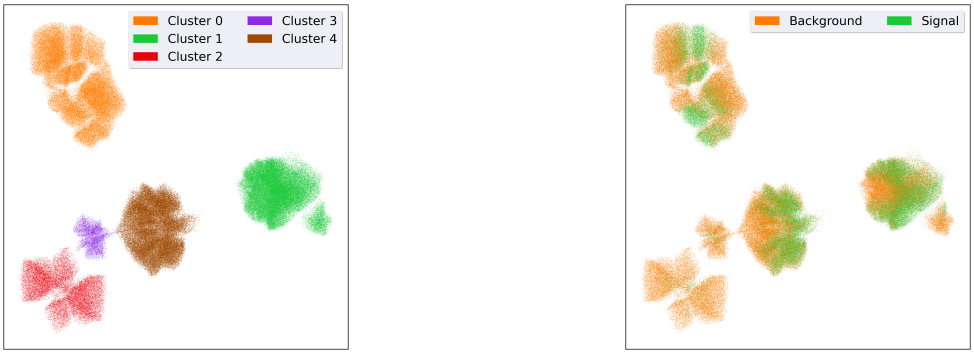
To calculate the AMS score, each prediction must be classified as a signal or background event. For this purpose, the predictions of the classifier were sorted according to their highest probability and only the highest N_{top} predictions were marked as signal events. A decision threshold is used to determine $N_{\text{top}} = \text{threshold} \times N_{\text{predictions}}$. This decision threshold was dynamically determined by a threshold scan using the predictions of the validation data. The decision threshold with the highest AMS score was then used to calculate the AMS score of the test data predictions.

[1] These weights come from the Monto Carlo simulation and are intended to compensate for the artificially enriched number of signal events.

[2] The HiggML training data has an Imbalance Ratio of $IR = \frac{N_{\text{background}}}{N_{\text{signal}}} = 1.92$

3.2. Two dimensional representation of the HiggsML dataset

The initial stage of the Federation training creates a low-dimensional representation of the training data. In order to enhance the UMAP embedding, we performed a grid search on the hyperparameter space on the validation set to determine the optimal values for the parameters `n_neighbors` and `m_dist`. No Imbalanced Learning techniques were applied during this search. We found that for `n_neighbors` = 50 and `m_dist` = 0.95 the Federation achieved the highest



(a) (b)

Figure 4: Two dimensional representations of the training data of the HiggsML dataset. Each point represents one high-dimensional datapoint and is colored according to its (a) Federation cluster/member index (b) truth label

AMS score on the validation data [14]. The resulting two-dimensional representation of the HiggML training data is shown in figure 4. Figure 4a illustrates that the Federation identified five global structures in the HiggsML data, leading to the creation of five FMs. A comparison between figure 4a and the signal and background distribution in figure 4b shows that cluster 0 and 3 contain as majority background events, leading to a class imbalance inside these clusters. This indicates already that Imbalanced Learning techniques are required.

3.3. Results

Table 1 presents the AMS scores obtained by the baseline methods defined in Section 3.1 and the Federation technique using two different types of classifiers. The hyperparameters from Section 3.2 were used for the Federation. In addition to the XGBoost baseline, an Artificial Neural Network[3] was trained to showcase the versatility of the Federation.

Training the common part of the Federation took approximately 20 minutes on a single node, while training each individual Federation member took up to 30 seconds[4]. The training duration for the single classifiers was up to 70 seconds. Hence, using the Federation technique for this dataset, the training time of the ML models, when utilizing multiple nodes, could be reduced by a factor of about 2.33 compared to using a singular ML model.

Several over- and undersampling techniques were investigated during our research [14]. `RWO_sampling` and `CNNTomekLinks` from [10] were used for the last Federation result in table 1. The cluster-based classifier baseline performs worse than the XGBoost and ANN single classifier baselines, implying that the physics-based clustering approach of forming an expert ensemble

[3] The ANN was trained with 5 layers, consisting of [64, 64, 64, 32, 8] nodes per layer with a dropout rate of 0.1 and L2 regularization of 0.0001, using the ReLU activation function and the Adam optimization algorithm. The evaluation metric used was the AUC-PR. Early stopping was employed during training. The batch size for training was set to 512.

[4] Using an AMD Ryzen 7 5800X CPU with 64 GB of RAM.

Table 1: Approximate Median Significance (AMS) calculated from prediction of the HiggsML test dataset for different machine learning methods. A bootstrap with $N = 100\,000$ bootstraps was performed to estimate the 2σ confidence interval.

	XGBoost		Neuronal Network	
Method	AMS	2σ-CI	AMS	2σ-CI
Single classifier baseline	3.627	$[3.557, 3.697]$	3.495	$[3.428, 3.562]$
Cluster-based baseline	3.425	$[3.301, 3.560]$	3.329	$[3.157, 3.324]$
Federation	3.557	$[3.485, 3.630]$	3.114	$[3.057, 3.172]$
Federation (with over- and undersampling)	3.685	$[3.595, 3.775]$	3.514	$[3.446, 3.582]$

is not effective. However, when the *FM*s use the XGBoost model, the Federation's clustering technique outperforms the cluster-based XGBoost baseline. This suggests that clustering with UMAP and HDBSCAN produces a more effective clustering of distinct local regions in the feature space of the data. On the other hand, the Federation which used ANNs produced an AMS score lower than the single classifier baseline. Nevertheless, the Federation surpassed both the XGBoost and ANN single classifier baselines if Imbalanced Learning techniques were applied during the training of its members. This indicates that the ANN model is more affected by high class imbalances in the *FM*s than XGBoost. The Kaggle team behind XGBoost during the HiggML challenge achieved with their untuned model[5] an AMS score of 3.64655 [13] which is comparable to our single classifier XGBoost baseline.

4. Summary and conclusion

The Federation is a new ML technique which was developed for tackling the issue of the increasing computational resource demands caused by ever growing dataset sizes in HEP.

Compared to a single classifier, the Federation technique can significantly reduce computational overhead by distributing training and inference of its members across multiple independent nodes, thereby managing large datasets more effectively and decreasing overall training time and computational resource requirements[6]. This advantage comes at the expense of a manageable overhead during training and inference due to the common part of the Federation (UMAP embedding and HDBSCAN clustering), which can be minimized by using a smaller training size for the UMAP embedding [14]. Utilizing the Federation technique becomes increasingly beneficial as the complexity of Federation member models rises.

Our research demonstrates that this new method can effectively analyze HEP data and is able to outperform single classifier baselines or a similar cluster-based approach.
The source code for this new machine learning technique is accessible on GitHub [15].

References

[1] C Adam-Bourdarios, G Cowan, C Germain, I Guyon, B Kégl and D Rousseau. The Higgs boson machine learning challenge. *Proc. of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, PMLR **42** 19–55. 12 2015

[2] L Maaten and G Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, JMLR **9** (86):2579–2605. 2008

[3] L McInnes, J Healy, N Saul, and L Grossberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* JOSS **3** (29):861. 2018 z

[5] Feature engineering was used in their model

[6] However, for smaller datasets and less complex models, as used in this study, the computational effort for the common part of the Federation may be greater than the computational resources required by the individual FMs.

[4] L Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*. **29** (6):141–2. 2012
[5] T Sainburg, L McInnes and T Gentner. Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning. *arXiv preprint* arXiv:2009.12981 2021
[6] L McInnes, J Healy and S Astels. hdbscan: Hierarchical density based clustering *Journal of Open Source Software* JOSS **2** (11) 2017
[7] T Akiba, S Sano, T Yanase, T Ohta and M Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv preprint* arXiv:1907.10902 2019
[8] D Moulavi, P Jaskowiak, R Campello, A Zimek and J Sander. Density-Based Clustering Validation. *Proc. of the 2014 SIAM International Conference on Data Mining* 839–847. 2014
[9] H He and E Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* **21** (9):1263–1284. 2009
[10] G Kovacs. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing* **83** 2019
[11] H He, and Y Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press 2013
[12] R Cruz, L Hafemann, R Sabourin and G Cavalcanti. DESlib: A Dynamic ensemble selection library in Python. *arXiv preprint* arXiv:1802.04967 2018
[13] T Chen, and T He. Higgs Boson Discovery with Boosted Trees. *Proc of the NIPS 2014 Workshop on High-energy Physics and Machine Learning* (pp. 69–80). 2015
[14] M Mucha. Tackling large and imbalanced data in high energy physics by using a federation of binary classifiers *Master thesis* University of Bonn, 2022
[15] M Mucha. Code repository `https://github.com/mjmucha/federation`, 2023