Filomat 32:5 (2018), 1711–1725
https://doi.org/10.2298/FIL1805711M

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Deep Extreme Feature Extraction: New MVA Method for Searching Particles in High Energy Physics

Chao Ma[a,e], Jinhui Xu[b], Tiancheng Hou[c], Bin Lan[d], Zhenhua Zhang[e,*]

[a] *University College London, Department of Computer Science*
[b] *Arizona State University, School of Mathematical & Statistical Sciences*
[c] *New York University, Department of Finance and Risk Engineering*
[d] *the City University of Hongkong, Department of Economics*
[e] *Guangdong University of Foreign Studies, Department of Mathematics*

Abstract. In this paper, we propose Deep Extreme Feature Extraction (DEFE), a new ensemble MVA method for searching $\tau^+\tau^-$ channel of Higgs bosons in high energy physics. DEFE can be viewed as a deep ensemble learning scheme that trains a strongly diverse set of neural feature learners without explicitly encouraging diversity and penalizing correlations, which is achieved by adopting an implicit neural controller (not involved in feedforward computation) that directly controls and distributes gradient flows from higher level deep prediction network. Such model-independent controller results in that every single local feature learned are used in the feature-to-output mapping stage, avoiding the blind averaging of features. DEFE makes the ensembles 'deep' in the sense that it allows deep post-process of these features that try to learn to select and abstract the ensemble of neural feature learners. Based the construction and approximation of the so-called extreme selection region, the DEFE model is able to be trained efficiently, and extract discriminative features from multiple angles and dimensions, hence the improvement of the selection region of searching new particles in HEP can be achieved. With the application of this model, a selection region full of signal processes can be obtained through the training of miniature collision events set. In comparison with the Classic Deep Neural Network, DEFE shows a state-of-the-art performance: the error rate has decreased by about 37%, the accuracy has broken through 90% for the first time, along with the discovery significance has reached a standard deviation of 6.0 σ. Experimental data shows that DEFE is able to train an ensemble of discriminative feature learners that boosts the overperformance of final prediction. Furthermore, among high-level features, there are still some important patterns that are unidentified by DNN and are independent of low-level features, while DEFE is able to identify these significant patterns more efficiently.

1. Introduction

Particle accelerators are among of the most important tools in high energy physics research. The collision of proton creates a lot of particles as well as a large number of the data resource, which lays a foundation for

2010 *Mathematics Subject Classification*. 97R40; 97R30.
Keywords. Deep learning; Feature learning; Ensemble learning; Higgs bosons.
Received: 11 October 2017; Accepted: 17 November 2017
Communicated by Hari M. Srivastava
Corresponding author: Zhenhua Zhang
Email addresses: chao.ma.16@ucl.ac.uk (Chao Ma), jinhuixu@asu.edu (Jinhui Xu), tianchenghou@nyu.edu (Tiancheng Hou), binlan2-c@my.cityu.edu.hk (Bin Lan), zhenhuazhang@gdufs.edu.cn (Zhenhua Zhang)

the application of statistics as well as MVA techniques. The discovery of new particles is closely related to optimization of selection zone as well as the classification of signal events and background events. Hence, an effective model of statistics and Machine Learning is playing an increasingly significant role in high energy physics[22, 23, 32, 33]. Likewise, the challenging data from HEP would facilitate the invention and application of the new model of Machine Learning. The research to be conducted by is an aspect of this two-sided promotion.

Higgs boson, whose existence was temporarily confirmed in 2013, is an elementary particle in the Standard Model of particle physics[13]. In order to affirm the coupling effect between Higgs and Fermion and finally to verify the Standard Model, the study of decay channel$\tau^+\tau^-$ through the large hadron collider (LHC) is of great significance[25]. However, Higgs boson is often buried by a large number of background events, which makes it hard to be detected. Recently, ATLAS has detected the evidence of the decay from Higgs boson to $\tau^+\tau^-$ channel by BDT(Boosted Decision Tree, one of the state-of-the-art machine learning techniques). Since the signals are relatively weak and are buried in background noises. Hence, the significance of the observed deviation from BOH (short for Background-Only Hypothesis) is only 4.1 sigma. Hence, it is demonstrating to develop more sophisticated MVA methods which are expected to have higher sensitivity to signal events.

Our research is based on several kinematic features(both low-level and high-level features) of final state productions of MC simulated events. Low-level features are physical quantities of decay production can be observed by detectors of LHC such as CMS. High-level features are derivatives of low-level features calculated by physicists. Identifying signal events (short for collision events created by the $\tau^+\tau^-$ decay of Higgs boson) as well as selection region (short for the corresponding region of the decision areas of signal events in feature space) with a relatively high statistics significance and accuracy rate from a large number of background events(short for non-Higgs-boson events), is a difficult issue due to the high dimensionality and imbalanced nature of the data. Therefore, the relevant analysis is often based on sophisticate MVA methods based on machine learning, such as Boosted Decision Tree and neural networks. In fact, the requirements of classifiers are becoming stricter in order to improve the searching efficiency of LHC searching for new particles as well as the confidence level. The result of recent research suggests that, even with the help of experienced physicists, traditional classifiers such as SVM, NN, Decision Tree, Ensemble Learning and so forth, fail to detect all the significant structures hidden in data. Extracting high-level features automatically, Deep Learning is regarded as one of the new approaches to break through this limitation and promote the development high energy physics.

2. Deep Learning and Related Works

As a new learning algorithm of Multilayer neural network, Deep Learning[4], has become a great interest in the field of machine learning research and achieved great success in various of tasks[7, 9, 15–17, 20, 27, 31]. Deep Learning is not only capable to automatically design more complicated, distinct and nonlinear features (called feature learning), but also mitigate the local extremum problem of classical training algorithm.

However, the application of deep learning to high energy physics hasn't been studied until recently. Baldi.P et al.,2014[3] initially applies the classical Deep Learning approach to the identification of the Higgs boson(the counter channel of bottom quark-anti bottom quark). The experiment result expresses that the nonlinear features designed by Deep Learning algorithm possess good prediction capability. Compared with the features designed by physicists(later referred to as 'high-level features'), these nonlinear features increase the performance index by eight percent and reach the expected discovery significance(EDS) with five sigmas. The result shows that deep neural network unearths some important features ignored by physicists without drawing support from physical expertise, which indicates that the superiority of Deep Learning approach can be fully applied to in the data analysis of Large Hadron Collider.

It is worth noting that, though the performance of deep learning approach outperforms the hand-designed features of physicists when using the deep neural network of low-level feature training, further experiment result clarifies that the addition of high-level features does not improve the classification performance of the deep neural network. This phenomenon is explained as 'the algorithms are automatically discovering the insight contained in the high-level features' in the original paper of Baldi.P et al.,2014[3].

However, in our research, we found it is not the case. In the following part we would come up with a new model to give a different explanation to the above-mentioned phenomenon, that deep neural networks actually fails to fully discover the insight contained in the high-level features or neither completely excavate low-level features, hence resulted in the equivalent performances with or without high-level features. Thus, there is still a long way to go in the aspect of feature extraction.

In addition, classical deep learning algorithm needs a large number of training samples, thus resulted in a considerable amount of training time(it often takes days to train). In spite of using millions of training samples, the final accuracy index of the research is still less than 0.9, which also reflects the inefficiency of classical deep learning algorithms. Therefore, in conclusion, the research of applying deep learning to the discovery of new particle is still in the beginning stage, it still has certain one-sidedness in the extraction of high-level features and the optimizing of selection field.

3. DEFE: the proposed method

3.1. Introduction

Based on the analysis above, our research is focused on $\tau^+\tau^-$ of the Higgs boson, and we propose a new MVA method– the Deep Extreme Feature Extraction(DEFE for short) model. The idea of the model is, instead of directly approximating the ideological selection region, we divide the sample-variable space under a supervised setting and train multiple SDAENN[30] as well as the so-called extreme selection region, using which as a bridge finally to approximate globally and optimize the selection region of the hadron signal events.

More specifically, we supervisedly operate the space partition of product space between feature space and sample space by a weak classifier and divide it into a number of overlapping subspaces(this process is called the discriminative partition), maintaining at the same time the ratio balance between the background events and signal events on each subspace. Based on this, we build an SDAENN for each subspace to process partial feature extraction. The resulting selection region is called extreme selection region. Finally, we take the union of the features over all subspaces and approximate extreme selection region globally by only a single terminal classifier, in order to achieve the goal of multi-perspective feature extraction and feature appreciation as well as covering the Higgs boson's selection region as much as possible. The resulting selection region is called the approximated extreme selection region. In some cases, the extracted features are further reduced by PCA to obtain linearly independent features. In a macro context, this model embeds several unsupervised feature extraction in a large-scale framework of supervised feature extraction, avoiding the blindness and locality of single unsupervised pre-training. Therefore, DEFE can be regarded as a new ensemble learning method, and it is a thorough ensemble learning rather than a voting based ensemble learning.

3.2. Problem Formulation

Let the set of simulated event to be $\mathcal{D} = \{(\mathbf{x}_1, y_1, w_1), ..., (\mathbf{x}_n, y_n, w_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$, d is the dimensionality of the input feature, $y_i \in \{s, b\}$is the label of each event, meaning signal and background respectively. $w_i \in \mathbb{R}^+$ is weight associated with each event, which is intended to adjust the bias derived from the fact that the proportion of signal event in simulation may not be identical to the real prior class probability. Let $\mathcal{S}$ to be the set containing all signal events, $\mathcal{B}$ be the set containing background event, n_s to be the number of signal events, and n_b to be the number background events.The weight of each event should satisfy:

$$\sum_{i\in\mathcal{S}} w_i = n_s, \sum_{i\in\mathcal{B}} w_i = n_b, \tag{1}$$

Given a classifier $g : \mathbb{R}^d \to \{s, b\}$, we call $\hat{G} = \{\mathbf{x} \in \mathbb{R}^d, g(\mathbf{x}) = s\}$ the approximate selection region of classifier g. Let $G = \{\mathbf{x}_i, y_i = s\}$, $G_T = G \bigcap \hat{G}$. Then $\hat{n}_s = \sum_{i\in G_T} w_i$ is an unbiased estimator of the expected number of signal events which is selected by the classifier.

Then, objective of the problem is now to maximize the approximate median significance (AMS)[1], which defined as:

$$AMS = \sqrt{2((n_s + n_b + b_{regular}) \ln(1 + \frac{n_s}{n_b + b_{regular}}) - n_s)} \tag{2}$$

To simplify the problem,in this paper weights are normalized to be uniformly distributed in $\mathcal{S}$ and $\mathcal{B}$ respectively, i.e.:

$$w_i = \begin{cases} \frac{n_s}{|\mathcal{S}|} & i \in \mathcal{S}, \\ \frac{n_b}{|\mathcal{B}|} & i \in \mathcal{B} \end{cases}$$

3.3. Extreme Feature Extraction As Ensemble Learning With Diversity

Before introducing the idea of Extreme Feature Extraction (EFE), we first briefly recap the formulation of ensemble learning that is closely related our proposed model here. Ensemble learning is an important strategy for improving the performance and accuracy of machine learning algorithms.Ensemble learning tries to learn a linear combination of base models of the following form:

$$f(y|\mathbf{x};\Theta) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} f(y|\mathbf{x},\Theta_m)$$

The key to the success of ensemble learning relies on the diversity of each base model $f(y|\mathbf{x};\Theta_m)$. If these base models are trained with decorrelated errors, their predictions can be averaged to improve performance. Thus, a set of classifiers (or experts) are trained to solve the same task under slightly modified settings (e.g., different batch of training examples, a different set of variables, or different random initializations). During the test period, predictions from multiple classifiers are then averaged to a final prediction that is expected to be more accurate and robust.

It's natural to improve the performance of deep learning by training an ensemble of neural networks with different initializations. Ensemble deep learning forms many state-of-the-art solutions of different large-scale tasks[24, 29]. However, in such vanilla ensemble learning, sub-neural networks are not trained with respect to a unified loss function (i.e., not ensemble-aware), and no efforts are made to improve diversity [18]. To overcome this, different schemes of explicitly encouraging diversity or penalizing correlations [2, 14, 19] are proposed. It's then trivial to generalize these models to the task of feature learning by training auto-encoders as base models. Nevertheless, these frameworks are not well-suited for feature learning tasks, since model averaging are often taken over final output rather than features learned by base models. Direct averaging over learned features might be unstable. Also, vanilla ensembles of feature learners are generally 'shallow' in the sense that base models are ensembled linearly, which might have an impact of pushing each base models towards the target too aggressively, resulting in a potential reduction of diversity.

Now we introduce an alternative scheme of performing ensemble feature learning, i.e. Extreme Feature Extraction (EFE). Let

$$\mathbf{H}_m(\mathbf{x};\Theta_m^f), m = 1, ..., |\mathcal{M}|$$

Be the set of neural feature mappings (which can be initialized by excatly the same initial parameters), where Θ_m^f is the parameters of the m^{th} feature map. Assume $\mathbf{H}$ be the matrix concatenating every sub feature matrix $\mathbf{H}_m$. Thus, In EFE, the model can be described by the following feature extraction - output model:

$$f(y|\mathbf{x};\Theta) = \mathbf{F}(\mathbf{H}(\mathbf{x};\Theta^f);\Theta^o)$$

Where $\mathbf{F}$ is the deep neural predictor that defines the feature-to-output mapping and Θ^o the corresponding parameters, and $\Theta_m^f = \{\Theta_m^f\}_{1 \le m \le |\mathcal{M}|}$. So far the structure of EFE bears no difference from classical deep

neural nets. The discriminating feature of EFE that forces each neural feature extractor to be diverse is the implicit neural controller (gating function) that is not involved in the feedforward compuation with $|\mathcal{M}|$ dimensional output defined by $\mathbf{g} := \mathbf{g}(\mathbf{x}; \Theta^g)$, which controls the gradient flow during learning:

$$\frac{\partial L_0(y, \mathbf{x}; \Theta)}{\partial \Theta_m^f} = \mathbf{g}_m(\mathbf{x}; \Theta^g) \frac{\partial L_0(y, \mathbf{x}; \Theta)}{\partial \mathbf{H}_m} \frac{\partial \mathbf{H}_m(\mathbf{x}; \Theta_m^f)}{\partial \Theta_m^f}$$

When these feature mappings are parameterized by deep neural networks, EFE model becomes Deep EFE (DEFE) model. The proposed EFE model has a number of desired properties. Firstly, the neural controller $\mathbf{g}$ directly distributes the gradient flows toward different feature learners, forcing thee learned features to be strongly diverse. Thus, EFE can be viewed as an ensemble learning scheme that only updates a small set of base feature learners by modifying the information of gradients, thus resulting in a diverse set of feature learners. Secondly, since EFE trains ensembles of feature learners without explicitly getting involved in the final averaging function, every single local feature learned are used in the feature-to-output mapping stage, avoiding the blind averaging of features. Thus, DEFE makes the ensemble 'deep' in the sense that it allows deep post-process of these features that try to learn to select and abstract the ensemble of neural feature learners. Thirdly, even the feature-to-output mapping $\mathbf{F}$ is set to be an averaging error, diversity is still not eliminated due to the implicit controller $\mathbf{g}$.

However, these advantages come with the difficulty of training the gating function $\mathbf{g}$ due to the fact that $\mathbf{g}$ itself is not incorporated into the loss function and network structure. In the following of the paper, we incorporate the gating function into the loss function by simple linear combination:

$$L_{EFE}(y, \mathbf{x}; \Theta) = \lambda \min(L_g(y, \mathbf{x}; \Theta^g), \delta) + L_0(y, \mathbf{x}; \Theta)$$

Where $L_g(y, \mathbf{x}; \Theta)$ is the loss function of training $\mathbf{g}$ toward target y. Through such incorporation of gating function into the total loss function, discriminative information from output targets are allowed to train the gating function. We restrict $|\mathcal{M}|$ to be even: when the dimension of y is not equal to $|\mathcal{M}|$, a binary tree of $\mathbf{g}$ (i.e., the discriminative partition to be introduced in the following of the paper) is trained to match the dimension of the target and minimize $\min(L_g(y, \mathbf{x}; \Theta), \delta)$. The reason that we employ $\min(\cdot, \delta)$ on $L_g(y, \mathbf{x}; \Theta)$ is to restrict the discriminative information from the targets $\mathbf{y}$, so that each feature learner are trained with approximately equal emphasis. Since training the model by a unified manner may be numerically stable and computationally expensive, in this paper, we introduce an algorithm in which $\mathbf{g}$, $\mathbf{H}$, and $\mathbf{F}$ are trained sequentially and greedily to obtain a good enough estimation of DEFE's parameters.

3.4. Constructing and Learning of the Extreme Selection Region

In this section, we introduce the formal description of the pratical algorithm that trains an DEFE ensemble. We first give a few definitions needed to describe the DEFE algorithm:

Definition 1. *Given a classifier $g : \mathbb{R}^d \to \{s, b\}$, we call $\hat{G} = \{\mathbf{x} \in \mathbb{R}^d, g(\mathbf{x}) = s\}$the approximate selection region of classifierg. Let $G = \{\mathbf{x}_i, y_i = s\}$, then $G_T = G \bigcap \hat{G}$ is called the hit selection region.*

Definition 2. *Given a classifier g, the approximate rejection region is defined as $\hat{H} = \{\mathbf{x}_i, g(\mathbf{x}_i) = b\}$. Let $H = \{\mathbf{x}_i, y_i = b\}$, then $H_T = H \bigcap \hat{H}$ is the hit rejection region.*

Definition 3. *Given the classifier g, we call $T = G_T \bigcup H_T$ the hit region, and $F = X \setminus T$ the anomalous region. Then, we can define the discriminative partition of the training example space as the tuples $\{T, F, \hat{G}, \hat{H}\}$.*

From the definition above, it's easy to see that the hit region and anomalous region is exactly the correctly classified and miss-classified samples, respectively. The reason that separate treatment of samples that counts for the fictitious knowledge (i.e., $\{\hat{G}, \hat{H}\}$) of the weak classifier is that we want to further characterize the decision boundary trained by a first and quick 'glance' at the data. We can further perform discriminative partition over the resulting regions $\{T, F, \hat{G}, \hat{H}\}$ respectively. By doing this procedure

recursively for n iterations, we can obtain 4^n partition of the sample space. In this paper, we consider the case that n is sufficiently small.

Hit region and anomalous region characterize the two different regions of the sample space that exhibit potentially different patterns and distributions of high-level features, therefore a single classifier might fail to capture such information. To balance the number of samples of the partition, we normally set classifier to be either a weak classifier (e.g. Decision trees) or a neural network that is not fully trained. Furthermore, 'weak' discriminative partition obtained via such weak classifier is, in fact, the decision boundary trained by a first and quick 'glance' at the data, thus information containing the partition of $\{\hat{G}, \hat{H}\}$ represents the subspace with principal different the structures hidden in the data. In contradiction to cluster analysis, discriminative partition tries to make use the information of the labels. The problems of overfitting might exist both due to the partition itself and the random errors from the weak classifier. To avoid this, we propose an additional procedure of random interchange, i.e. randomly select the samples from both hit region and anomalous region according to a preset ratio and switch these selected samples. This additional procedure will not only balance the partition but also enhance the robustness.

Now, we consider the partition against the feature space, i.e. the set containing every input attributes. In our work, we partition the feature set according to its physical interpretations. Note that overlapping of the partition is allowed. Given the partition $S = \bigcup S_i$, we are now able to define the following procedures.

Definition 4. *Let $X = \bigcup_{i=1}^{4^n} X_i$ be a discriminative partition of the sample space, and $S = \bigcup_{i=1}^{m} S_i$ a given partition of the feature space; Then we call $X \otimes F = (\bigcup X_i) \otimes (\bigcup S_i)$ a partition of the sample-feature space. Every resulting subsets forms a new set of $U = \{X_i\} \otimes \{S_j\} = \{(X_i, S_j)\}$, where $\otimes$ is the direct product.*

Definition 5. *From very subset $D_h \in U = \{X_i\} \otimes \{S_j\}, h = 1, ..., 4^n \times m$ of the sample-feature space, we choose/train the corresponding classifier g_h and its approximate selection region $\hat{G}_h$. Then, we define $\hat{G}_E = \bigcup_h \hat{G}_h$, as the extreme selection region. Similarly, we can define as the extreme hit region $G_{ET} = G \bigcap \hat{G}_E$. The process of generating and constructing the extreme hit region based on the classifier chosen is called the expansion of the selection region. Similarly we can define the process of the expansion of H.*

It's trivial to see that the process of expanding selection region always increases the number of samples that can be possibly covered by a set of multiple classifiers, i.e. $G_T \subset G_{ET}$. However, one primal concern might be that since discriminative partition and expansion of selection region closely rely on the label of the data, how can one guarantee that the selection region is still expanded without the prior knowledge of labels of the testing data? The key fact to solve this question lies in the fact that apart from the training data (including labels), the definition of selection region only depends of the resulting decision boundaries that can be well described and parameterized by classifiers g and g_h(even with simple rules in the case of decision tree based discriminative partitions). As a result, information regarding these regions are compressed by a limited number of classifiers rather than the raw sample-feature space $X \otimes S = (\bigcup X_i) \otimes (\bigcup S_i)$. Thus, although the previously described expansion of selection region technique cannot be directly used for deriving a divide-and-conquer mixture of classifier model, with the help of the resulting selection regions as stepping stones, 'extreme' information can then be unfolded and approximated by a single strong classifier.

In conclusion, the problem of improving the performance of deep learning can now be converted to the problem of approximating the expanded the selection region by merely a single classifier. In previous work of ensemble learning [5, 6, 10, 11, 26] tries to unify every sub-classifier g_h by an ensemble procedure of linear weighting, voting or winner-take-all, and achieves a fairly good result compared to a single classifier. As stated above, nevertheless, in the task of recognition of Higgs Bosons, this class of ensemble algorithms (Boosted Decision Tree for example) failed to significantly improve the performance of classification. The reason might be two folds: firstly, when applying the divide-and-conquer principle to the sample-feature space, only the shallow and presentational are exploited, thus missing local high-level information; secondly, only the weak classifiers' final output is considered, therefore in intrinsic structures and learned representational features are ignored. Also, it's too computationally expensive to apply directly ensemble learning algorithms to deep learning algorithms.

3.5. Greedy Training Algorithm for DEFE

To contribute to overcoming these difficulties, we have introduced the idea of feature learning from Deep Learning framework, and propose a new algorithm, the Deep Extreme Feature Extraction (DEFE). Now, we introduce a greedy training algorithm for DEFE. As depicted in Figure 1, in our prototype DEFE algorithm, the initial controller g is chosen to be a neural network or decision tree, and $\mathbf{H}_h$ to be the Stacked Denoising Autoencoder Neural Networks (SDANN). In this setting, DEFE is not allowed to utilize the output of each classifier; instead, the union of all the high-level features (the output of the final hidden layer) learned by each SDANN (i.e., the feature set of $\hat{G}_E$). Based on this feature set, a final deep neural predictor is employed to reorganize the extreme feature set, and learn to approximate the extreme selection region $\hat{G}_E$. By establishing this framework, both advantages of prior experiences of extreme selection region and the feature extraction power of deep learning techniques are combined. The local features on $\hat{G}_E$ are thus reorganized into high-level features learned by the final deep classifier. With the existing mature training algorithms of deep learning to train the final deep classifier, the expensive computational cost of apply ensemble learning directly to learning the gating weights of each classifier g_h can be also avoided. The DEFE algorithm applied to the optimization of recognizing Higgs Bosons are described as follows:

Input: the sample-feature space $X \otimes S$, labels $\{y_i\}$, and interchange rate α. We assume $n = 1$ and $m = 1$.

Step 1. *(Discriminative Partition): Train a neural controller on X, and obtain a partition of* $\{T, F, \hat{G}, \hat{H}\}$.

Step 2. *(Random Interchange) According to an interchange rate* α*, randomly exchange the elements between F and* T*,* $\hat{G}$ *and* $\hat{H}$*, respectively.*

Step 3. *(Partition of feature set): Given the feature set S, we deploy an overlapping partition. In the task of LHC hadron collisions, feature sets are partitioned as* $S = S_1 \bigcup S_2 \bigcup S_3$*, where* S_1 *is the momentum features,* S_2 *is the derivative of physical attributes, and* $S_3 = S$ *is the entire feature set.*

Step 4. *(The construction of extreme selection region): So far we obtained a partition U of the sample-feature space* $X \otimes S$*. For every* U_h,$h = 1, 2, ..., 4^n \times m$*, we train an SDAENN, denoted as* $\mathbf{H}_h$*. Note that the number of units in the first layer far outnumbers the length of the input vector, and the number of hidden units at each layer decreases gradually to a fix number K as depth increases to compress the information. In order to make every SDAENN equally important,K is fixed as 50. All SDAENNs are trained unsupervised in order to learn non-trivial features (or optionally followed by supervised finetuning step with very few epochs). By training these* $4^n \times m$ *SDAENNs, we obtained implicitly the extreme selection region* $\hat{G}_E$*.*

Step 5. *(Combining extreme feature set): For every* $\mathbf{H}_h$*, we take their output* $S_h = \{S_{h1}, S_{h2}, ..., S_{hK}, \}$ *of the last hidden layers. Then, the extreme feature set can be constructed as* $S_E = \bigcup_h S_h$*, and the new sample-feature space becomes* $X \otimes S_E$*.*

Step 6. *(Learning and approximating* $\hat{G}_E$*): Finally, we train an deep neural network* $\mathbf{F}$ *on* $X \otimes S_E$ *as a final classifier with stochastic gradient descent. The resulting decision boundary will be a improved estimation of* $\hat{G}_E$*.*

4. Experiment

4.1. Methodology

Based on the simulated data, the proposed Deep Extreme Feature Extraction (DEFE) is used to learn the selection region (or extreme selection region $\hat{G}_E$). The goodness of such approximation is usually measured by various metrics. In this paper, The metric used for the goodness of fit comparison is the total area under the Receiver Operating Characteristic curve (ROC), i.e. The AUC metric. In general, a higher value of AUC represents higher classification accuracy averaged across a wide range of different choices of thresholds. The expected significance of a discovery (in units of sigmas) is also calculated for 100 signal events and 1,000 background events. It denotes the significance of null selection region hypothesis (or the discovery

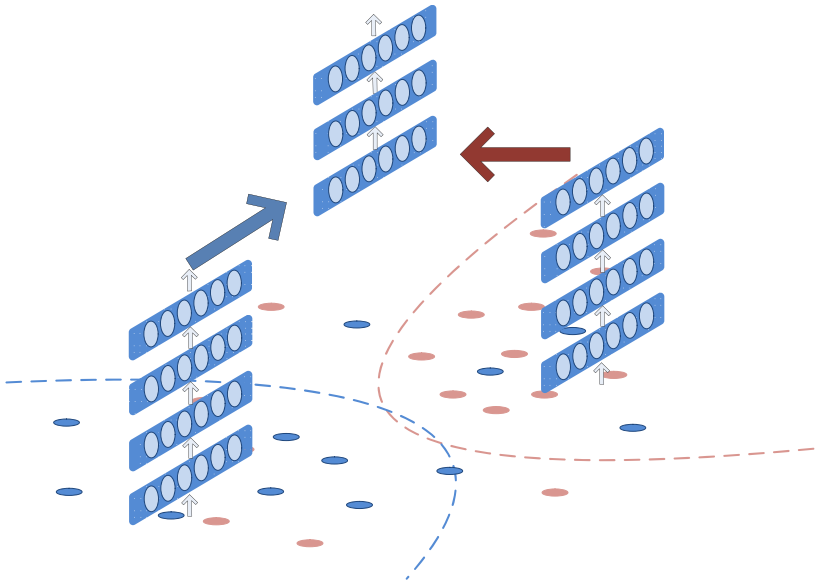

Figure 1: How greedy DEFE algorithm is constructed. Two dashed lines with different colors illustrate one possible partition defined by decision boundary of the controller g. Deep feature extractors are then built, receiving gradient flows from each interior of decision boundary. Note that this is not a divide-and-conquer algorithm: g is model-independent in the sense that it is not involved in feedforward computation. For every new test example, features from all deep feature extractors are computed simultaneously, weighted equally, and forwarded to a final deep neural network that tries to learn to select useful locally significant features by approximating the extreme selection region, $\hat{G}_E$.

significance)[8]. If the resulting P-value of null selection hypothesis is below a certain value (normally required to be one millionth or lower, corresponding to discovery significance greater than 5-sigma), then the declaration of a new physics can be made. Once the selection region been trained, the model is ready for analyzing real experimental data.

4.2. Data

The data we use in our experiment is obtained from the Higgs Boson Machine Learning Challenge(data can be downloaded at http://www.kaggle.com/c/higgs-boson). Data is generated by an official simulator of ATLAS, with Higgs to Tau-Tau events mixed with different backgrounds. Based on current knowledge of particle physics, random collisions are simulated, tracked and detected by a simulated detector. The mass of the Higgs Boson is fixed at 125GeV, considering the following collision event:

1. Signal Event: The Higgs boson decays into $\tau^+\tau^-$.
2. Background Event 1: The Z bosons (91.2 GeV) decay into $\tau^+\tau^-$, which is similar to the signal event and becoming the difficult point in classification.
3. Background Event 2: A pair of top quarks is involved, accompany with a lepton and hadronic decayed τ.
4. Background Event 3: W bosons decay into an electron or a muon and a hadronic decayed tau.

The total number of events is 250,000. For any given collision event, the following 30 input attributes are obtained, with 17 low-level features measured by the detector and 13 additional high-level features calculated from low-level features, see Table 1.

4.3. Parameters and Training Strategy

We use a hundred thousand samples to train the DEFE model, and use about eighty thousand samples to test the DEFE model. ROC (Receiver Operating Characteristic Curve) is used to visualize the performance

Categories	High-level	Leptons	Hadronic Tau	Jets	Neutrinos
Variables	13 high-level features in total.	1,Transverse momentum 2,Azimuth angle 3,Pseudorapidity	1,Transverse momentum 2,Azimuth angle 3,Pseudorapidity	1,Number of jets 2,Transverse momentum of the leading jet 3,Azimuth angle of the leading jet 4,Corresponding features of the subleading jet; 5,Total transverse energy 6,And more, see Appendix B	1,Missing transverse momentum 2,Azimuth angle 3,Total transverse energy

Table 1: Kinematic Features

and. The AUC (Area under the Curve of ROC) and Expected Discovery Significance are used to quantify the performance.

All data are normalized. Afterwards, we do an $n = 1$ discriminative partition, on each subset, with random swap ratio α=0.05. In other words, we partition the original dataset into four overlapped subsets. Finally, we employ SDAENN to gain the high-level feature on a $m = 3$ partitioned feature space, on each of the data subset, gaining altogether twelve high-level feature sets. The SDAENNs are chosen to have fifty output unit, so by complying the steps above, we can ultimately obtain the so-called "extreme features" with 12×50=600 dimensions. And then, before inputting into the DNN classifier, we reduce the dimension to 300 by PCA.

In our model, each of the SDAENN on their corresponding sample-feature partition is set to have the following parameters: For all SDAENN: Totally five hidden layers, the output layer has fifty neural units. For each feature space, the structure of each hidden layer is given as S_1 : {250,200,150,100,50}; S_2: {200,200,150,100,50}; S_3: {300,250,200,200,50}.

Among our experiments, the activation function is set to be a sigmoid function. One can also implement Rectified Linear Units (ReLU) [21] instead of sigmoid function. In our case, however, since we will be interested in visualizing features (i.e., outputs of nonlinear activation function) learned by our algorithm, it is more convenient to use a sigmoid function that can automatically squash their outputs between 0 and 1. Therefore, we avoid using ReLU activation function due to the fact that they tend to blow up their activations (i.e. the range of output is not constrained for ReLU). The training algorithm is plain stochastic gradient descent (SGD) with batch training and momentum. Batch size of SGD is one hundred, the momentum is 0.5, and learning ratio is 0.1 in the beginning and decrease in the training process, the descending ratio is 0.997. Under the fine-tuning phase, we adopt the following early-stop strategy: Stop training if cross-validation error of SDAENN increase to 0.002 above the minimum, or the change of cost is lower than 0.0001 after 10 iterations. Under this strategy, the fine-tuning normally stop after 70 120 iteration. This effectively deterred over-fitting. For each neural feature learner, we adopt an additional supervised fine-tuning step with only 10 epochs. The parameters stated above are also used in the terminal classifier (DNN). The popular drop-out training technique [28] is not used because of their deterioration on accuracy was found in our preliminary experiments, possibly due to its side effect of introducing noise [12] outweighs its positive effect of automatically implementing variational Bayesian inference in our case.

5. Results

Table 1 demonstrates the collection of the thirty-dimensional feature used in our model. In Table 2 we observe the comparison of AUC accuracy rate between DEFE model and other baseline models. Among which, the training sets contain 80,000 samples, and if not specially addressed, low-level features and high-level features are all adopted(if not adopt high-level features, then the performance of DEFE and DNN are much equivalent). The expected significance of a discovery (in units of Gaussians) for 100 signal events and 1,000 background events. The calculation of expected statistical significance is referred to the method

Model	AUC	Discovery Significance:Z
DEFE	**0.916**	**6.0**σ
DEFE(low features only)	0.898	5.6σ
DNN(low features only)	0.880	4.9σ
DNN	0.885	5.0σ
SVM	0.76	3.5σ
NN	0.81	3.7σ
Boosted Decision Tree	0.816	3.7σ
Random Forests(RF)	0.84	3.9σ

Table 2: Algorithm Comparison

presented in document [3]. In[3], a slightly different task that the case of a pair of leptonic decay of Taus is considered. Due to the similarities of both events and features, their results are also listed for comparison.

Compared with classic Deep Neural Network (DNN) under the restriction of 90% background rejection rate, the error rate of DEFE drops by approximately 37%, and the precision indicator of AUC breaks through 90% for the first time, with statistical significance reaching as high as 6.0 σ. It is also worth noting that, unlike DNN, the additional high-level features promote the accuracy of DEFE significantly. In other words, DEFE can learn essential features more effectively from additional high-level features.

Finally, it's worth mentioning that DEFE does capture some important features of $Higgs \rightarrow \tau^+\tau^-$ channel. Appendix I illustrates first 30 of the features extracted by DEFE. Obviously, automatically learned features by DEFE exploit to the full the discriminative power hidden under raw input features. Note the great diversity among different feature learners trained by DEFE algorithm. Among high-level features, there are still some important patterns that are unidentified by DNN and are independent of low-level features, therefore the DNN's treatment of high-level and low-level features are insufficient, while DEFE is able to identify these significant patterns more efficiently. With the state-of-the-art performances of the proposed method, we hope to improve the analyzing quality of HEP data and the statistical significance of confirming the physical facts.

6. Conclusion

In this paper, we proposed a novel ensemble deep learning technique, Deep Extreme Feature Extraction (DEFE), to the task of identifying Higgs Bosons(Tau-Tau channel) from background signal. Based the construction and approximation of the so-called extreme selection region, the model is able to efficiently extract discriminative features from multiple angles and dimensions and therefore boost the overall performance. The result is improved in approximately one σ compared to DNN. In comparison with the traditional deep learning algorithm, we discover that performance of DEFE is significantly boosted with high-level feature inputs, avoiding the equivalent performances with or without high-level features. This result indicates that unlike the vanilla deep neural network, DEFE successfully trains a diverse set of neural feature learners, and discover the excess discriminative information contained in the high-level features. In the future, it's still an open question to propose further training algorithms to train an EFE model universally and efficiently. [12, 21, 28]

Acknowledgments

We would like to thank Dr. Yanjun Tu from Department of Physics, the University of Hong Kong for her valuable discussions at the early stage of this work. We also thank MRIIS members Longxin Li, Qing Wen, Jiangnan Huang, Ziyu Lin, Yingli Wang for their kind support. This paper is funded by the National Natural Science Foundation of China (No. 71271061), National statistical research key projects (No.2016LZ18),

Philosophy & Social Science Project (No. GD12XGL14) & Soft Science Project (No. 2015A070704051) & Natural Science Projects (No. 2014A030313575, 2016A030313688) & Quality engineering and teaching reform project (No.125-XCQ16268) of Guangdong Province, Science & Technology Fund of Guangdong Education Department (No. 2013KJCX0072), Philosophy & Social Science Project of Guangzhou (No. 14G41), Special Innovative Project (No. 15T21) & Key Team (No. TD1605) & Major Education Foundation (No. GYJYZDA14002) & Higher Education Foundation (No. 2016GDJYYJZD004) of Guangdong University of Foreign Studies, Climbing Plan Foundations of Guangdong (No. pdjh2015a0180, pdjh2016a0166). Chao Ma would like to thank CSC Foundation for supporting his postgraduate research.

References

[1] Claire Adam-Bourdarios, Glen Cowan, Cecile Germain, Isabelle Guyon, Balzs Kgl, and David Rousseau. Learning to discover: the higgs boson machine learning challenge. *Machine Learning*, 2014.
[2] Monther Alhamdoosh and Dianhui Wang. Fast decorrelated neural network ensembles with random weights. *Information Sciences*, 264(6):104–117, 2014.
[3] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5(5):4308–4308, 2014.
[4] Yoshua Bengio. Learning deep architectures for ai. *Foundations & Trends in Machine Learning*, 2(1):1–127, 2009.
[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
[6] Leo Breiman and Leo Breiman. Bagging predictors. In *Machine Learning*, pages 123–140, 1996.
[7] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pages 160–167, 2008.
[8] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *European Physical Journal C*, 71(2):1–19, 2010.
[9] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *University of California Berkeley Brigham Young University*, pages 647–655, 2013.
[10] Yoav Freund, Yishay Mansour, and Robert E. Schapire. Generalization bounds for averaged classifiers. *Annals of Statistics*, 32(4):1698–1722, 2004.
[11] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer & System Sciences*, 69(7):15421545, 1986.
[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.
[13] David Griffiths and Gerald W. Intemann. Postuse review: Introduction to elementary particles. *American Journal of Physics*, 58(3):282–283, 1990.
[14] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Nips*, pages 1799–1807, 2012.
[15] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, and Adam Coates. Deep speech: Scaling up end-to-end speech recognition. *Eprint Arxiv*, 2014.
[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25(2):2012, 2012.
[17] Quoc V. Le, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Marc'Aurelio Ranzato, Jeffrey Dean, and Andrew Y. Ng. Building high-level features using large scale unsupervised learning. *CoRR*, abs/1112.6209, 2011.
[18] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David J. Crandall, and Dhruv Batra. Why M heads are better than one: Training a diverse ensemble of deep networks. *CoRR*, abs/1511.06314, 2015.
[19] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Networks the Official Journal of the International Neural Network Society*, 12(10):1399–1404, 1999.
[20] A. Mohamed, G. E. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio Speech & Language Processing*, 20(1):14–22, 2012.
[21] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
[22] D. C. O'Neil and Atlas Collaboration. Tau identification using multivariate techniques in atlas. In *Journal of Physics Conference Series*, pages 597–604, 2008.
[23] Byron P. Roe, Hai Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon Mcgregor. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments & Methods in Physics Research*, 543(2-3):577584, 2004.
[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
[25] Nils Ruthmann. Evidence for higgs boson decays to the $\tau^+\tau^-$ final state with the atlas detector. *8000 GeV-cms*, 2013.
[26] Robert E. Schapire. *Theoretical Views of Boosting and Applications*. Springer Berlin Heidelberg, 2001.
[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Eprint Arxiv*, 2014.

[28] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
[29] C. Szegedy, Wei Liu, Yangqing Jia, and P. Sermanet. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.
[30] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.
[31] Mnih Volodymyr, Kavukcuoglu Koray, Silver David, Andrei A Rusu, Veness Joel, Marc G Bellemare, Graves Alex, Riedmiller Martin, Andreas K Fidjeland, and Ostrovski Georg. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–33, 2015.
[32] Shimon Whiteson and Daniel Whiteson. Machine learning for event selection in high energy physics. *Engineering Applications of Artificial Intelligence*, 22(8):12031217, 2009.
[33] Hai Jun Yang, Byron P. Roe, and Ji Zhu. Studies of boosted decision trees for miniboone particle identification. *Nuclear Instruments & Methods in Physics Research*, 555(1-2):370–385, 2005.

Appendix A: Visualization of Base Neural Feature Learners:

We present selected first 30 of the 600 features learned by 12 base feature learners. Note the great diversity among different feature learners trained by DEFE algorithm.

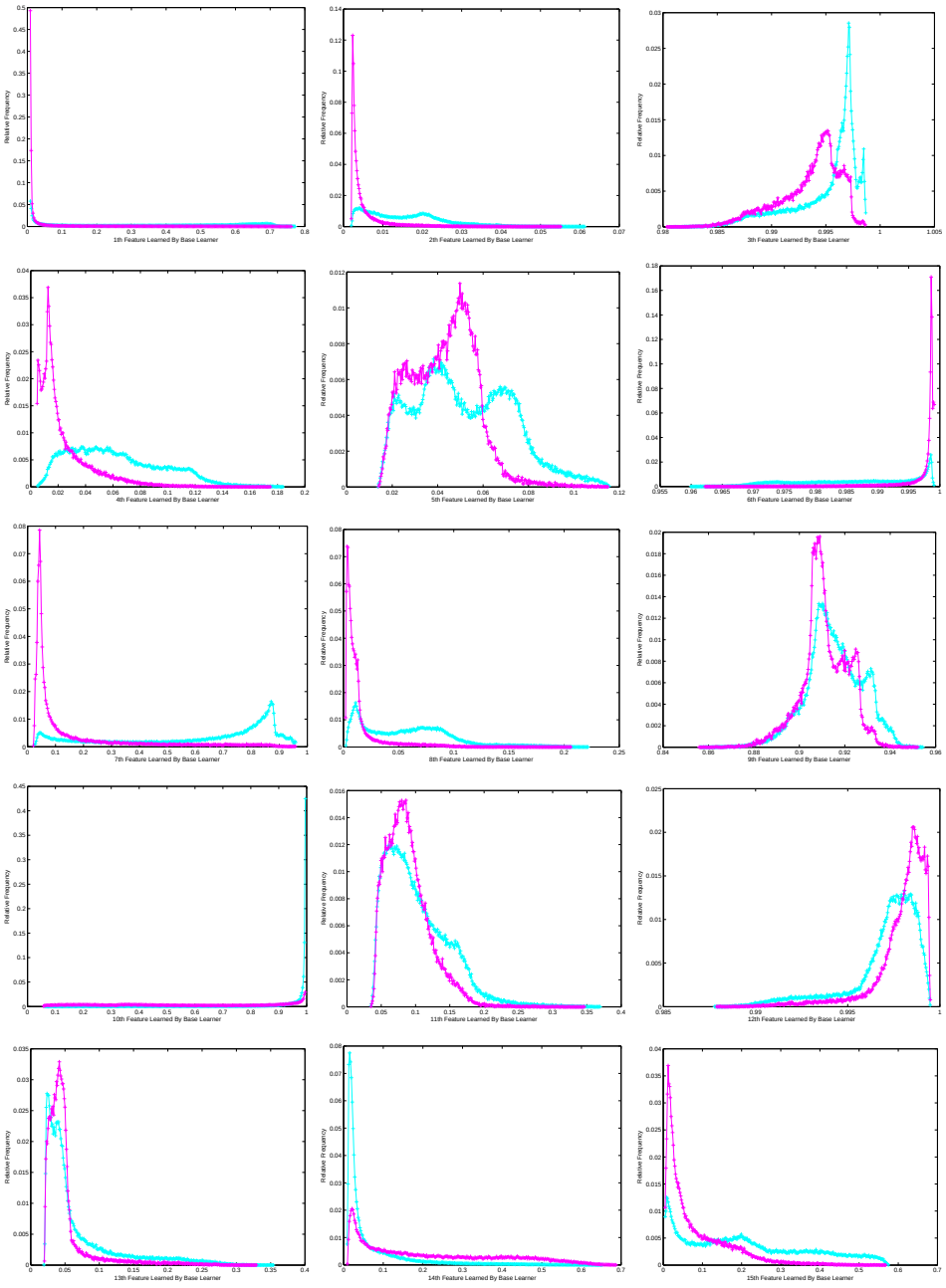

Figure 2: Relative fequency of features learned by feature learners, 1-15. Shimmering blue lines refer to signal events, while pink lines represent background signals.

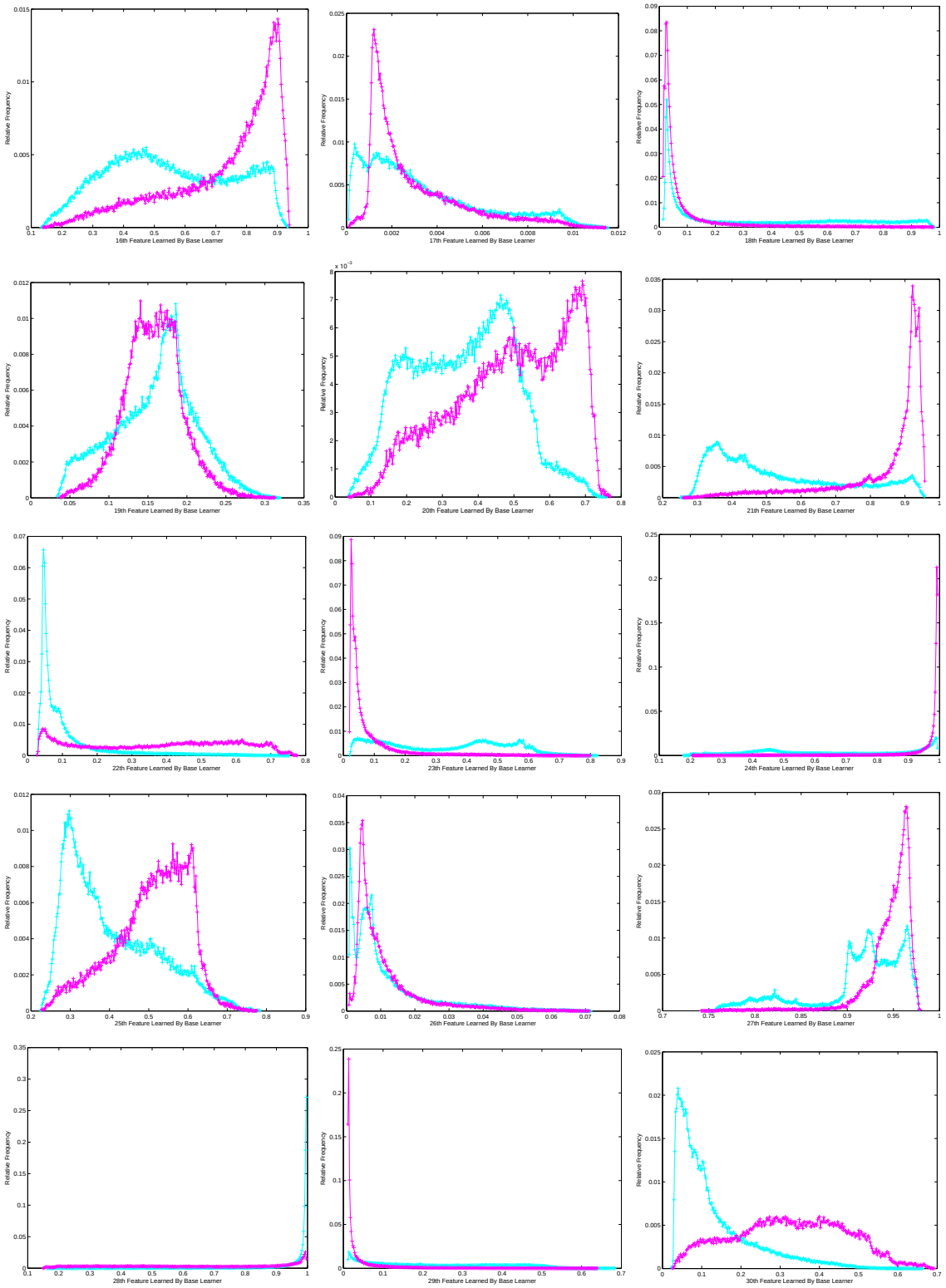

Figure 3: Relative fequency of features learned by feature learners, 16-30. Shimmering blue lines refer to signal events, while pink lines represent background signals.

Appendix B: Definition of Input variables[1]:

1. **DER_mass_MMC**: The estimated mass mH of the Higgs boson candidate, obtained through a probabilistic phase space integration.
2. **DER_mass_transverse_met_lep**: The transverse mass between the missing transverse energy and the lepton.
3. **DER_mass_vis**: The invariant mass of the hadronic tau and the lepton.
4. **DER_pt_h**: The modulus of the vector sum of the transverse momentum of the hadronic tau, the lepton, and the missing transverse energy vector.
5. **DER_deltaeta_jet_jet**: The absolute value of the pseudorapidity separation between the two jets (undefined if PRI_jet_num ≤ 1).
6. **DER_mass_jet_jet**: The invariant mass of the two jets (undefined if PRI_jet_num ≤ 1).
7. **DER_prodeta_jet_jet**: The product of the pseudorapidities of the two jets (undefined if PRI_jet_num ≤ 1).
8. **DER_deltar_tau_lep**: The R separation between the hadronic tau and the lepton.
9. **DER_pt_tot**: The modulus of the vector sum of the missing transverse momenta and the transverse momenta of the hadronic tau, the lepton, the leading jet and the subleading jet (if PRI jet num = 2) (but not of any additional jets).
10. **DER_sum_pt**: The sum of the moduli of the transverse momenta of the hadronic tau, the lepton, the leading jet and the subleading jet (if PRI jet num = 2) and the other jets (if PRI jet num = 3).
11. **DER_pt_ratio_lep_tau**: The ratio of the transverse momenta of the lepton and the hadronic tau.
12. **DER_met_phi_centrality**: The centrality of the azimuthal angle of the missing transverse energy vector w.r.t. the hadronic tau and the lepton.
13. **DER_lep_eta_centrality**: The centrality of the pseudorapidity of the lepton w.r.t. the two jets (undefined if PRI_jet_num ≤ 1).
14. **PRI_tau_pt**: The transverse momentum of the hadronic tau.
15. **PRI_tau_eta**: The pseudorapidity of the hadronic tau.
16. **PRI_tau_phi**: The azimuth angle of the hadronic tau.
17. **PRI_lep_pt**: The transverse momentum of the lepton (electron or muon).
18. **PRI_lep_eta**: The pseudorapidity of the lepton.
19. **PRI_lep_phi**: The azimuth angle of the lepton.
20. **PRI_met**: The missing transverse energy.
21. **PRI_met_phi**: The azimuth angle of the missing transverse energy
22. **PRI_met_sumet**: The total transverse energy in the detector.
23. **PRI_jet_num**: The number of jets (integer with the value of 0, 1, 2 or 3; possible larger values have been capped at 3).
24. **PRI_jet_leading_pt**: The transverse momentum of the leading jet, that is the jet with largest transverse momentum (undefined if PRI_jet_num = 0).
25. **PRI_jet_leading_eta**: The pseudorapidity of the leading jet (undefined if PRI jet num = 0).
26. **PRI_jet_leading_phi**: The azimuth angle of the leading jet (undefined if PRI jet num = 0).
27. **PRI_jet_subleading_pt**: The transverse momentum of the leading jet, that is, the jet with second largest transverse momentum (undefined if PRI_jet_num ≤ 1).
28. **PRI_jet_subleading_eta**: The pseudorapidity of the subleading jet (undefined if PRI_jet_num ≤ 1).
29. **PRI_jet_subleading_phi**: The azimuth angle of the subleading jet (undefined if PRI_jet_num ≤ 1).
30. **PRI_jet_all_pt**: The scalar sum of the transverse momentum of all the jets of the events.