University of North Carolina Asheville
Journal of Undergraduate Research
Asheville, North Carolina
May 2020

Application of Semi-Supervised Machine Learning Methods Towards the Identification of the Higgs Boson

Caleb Spradlin
Department of Computer Science
University of North Carolina at Asheville
Asheville, North Carolina

Faculty Advisor: Dr. Kevin Sanft

Abstract

Colliders allow physicists to probe the previously unknown world of sub-atomic physics employing observations of exotic particles through high-energy collisions. Physics communities regularly rely on hand-crafted, high-level features in conjunction with shallow machine-learning packages to accurately identify particles produced in collisions. This process proves excessive and time-consuming. This work provides an innovative means of solving the problem of accurate identification of Higgs boson particles through state-of-the-art, semi-supervised learning methods, and data collected by the European Centre for Nuclear Research. This research demonstrates how using semi-supervised learning techniques, specifically weight-averaged consistencies and data abstraction methods, alleviates the need for fully labeled datasets in accurate identification. Furthermore, it is demonstrated how deep semi-supervised learning models automatically extrapolate high-level features from the data given.

1 Introduction

There is a wealth of data that comes with most aspects of high-energy physics (HEP). The primary tools for investigating HEP are accelerators. These machines collide various fundamental particles to create strange and curious resulting particles that can only be studied at high-energy densities. The observation of these particles and the ability to measure them provide an exciting view of subatomic physics which may yield critical insights into the fundamental nature of matter. An abundance of statistical analysis in partnership with the field of HEP has been improved by recent advances in machine learning[1].

Despite this improvement, there is still much effort focused on the high-level feature engineering used in these machine learning algorithms. Teams of physicists spend a great deal of effort on hand-crafting specific field-related features that will benefit the machine learning model's classification abilities. While helpful, this approach can be labor-intensive and not always optimal. A deep neural network approach could obviate the need for manual feature extraction. Instead of manual derivation, the deep neural networks are capable of deriving high-level features automatically[2].

The Higgs boson was first observed in 2011 and has been the center of focus for the Large Hadron Collider (LHC). The Higgs field is a field thought to permeate the entire universe[3]. Without the Higgs field, subatomic particles would have no mass. Without such mass they would not be able to attract each other, leaving particles to float around at the speed of light. The more that physicists observe and study the Higgs boson, the more is known about how nature works at its most fundamental level

This research exploits the abilities of semi-supervised machine learning to produce an accurate, robust neural network classifier that can classify an event which produced a Higgs boson as an intermediary particle (signal) from an event which looks similar but did not produce a Higgs boson (background). This neural network classifier will be able to accurately classify such events without the need for a training dataset[4] containing complete, labeled examples. In addition, deep semi-supervised learning (SSL) techniques can manually extrapolate high-level features from low-level data gathered from runs of the LHC with little to no human interference. More importantly, SSL methods allow accurate models to be trained with less than half of the labeled data previously used[5]. This classification algorithm provides an excellent test case for innovative semi-supervised learning research and its application to real-world data that contains vast amounts of noise and nuanced features. Other existing approaches use classical, supervised deep learning techniques[6], which may be more susceptible to noise and misclassification in addition for the need of fully labeled data. Other solutions pre-process the labeled data into heat images of jet stream information collected by the detectors[7]. While this approach yields higher classification accuracy over classical deep learning techniques, it requires significant effort and computation time to pre-process large amounts of collider runs to create the necessary jet stream heat images.

Thus, while a small set of approaches have been previously explored, a semi-supervised learning application for this problem is still lacking. It is in this gap that the following is addressed:

- An innovative application of SSL is applied to high-energy physics in order to help identify unstable particles through inference and generative models
- It is shown how SSL can reduce the amount of labeled data that is necessary in a dataset without losing accuracy.
- It is demonstrated how deep semi-supervised learning classification tasks concerning colliders and high-energy physics can obviate the need for manual creation of high-level features currently used in classification tasks

2 Background

2.1 The Underlying Physics

Particle physics makes up the heart of the understanding of nature and the laws by which it abides. The standard model offers an innovative, unifying picture concerning the fundamental particles of the universe and the forces which control the interactions between such particles. All matter in the known universe consists of these fundamental particles, which occur in two types called quarks and leptons. In addition to providing an exciting view into fundamental matter, the standard model also describes a structure where the forces by which particles interact are mediated themselves by the exchange of particles.

Three of the four fundamental forces: the electromagnetic, weak nuclear, and strong nuclear force are all described by a Quantum Field Theory (QFT) for their respective fields. Each QFT for a certain force corresponds to the exchange of a force-carrying particle known as a gauge boson. The force-carrying particle for the electromagnetic field is the well-known photon. The gauge boson for the strong nuclear force is the gluon. And lastly, the force-carrying particles described by the weak nuclear force's QFT are the W and Z bosons.

While this theory is largely successful, there was a caveat in the equation. The W and Z bosons theoretically emerge without any mass, despite physicists having an abundance of experimental evidence suggesting that the W and Z bosons have a mass of 80.379±0.012 GeV/c^2, and 91.1876±0.0021 GeV/c^2 respectively. Thus, the final field and particle is the Higgs boson, discovered by the Compact Muon Solenoid (CMS) and A Toroidal LHC ApparatuS (ATLAS) experiments at the European Centre for Nuclear Research (CERN) in the Large Hadron Collider. This Higgs boson plays an integral role in the workings of the standard model; it provides a structure from which all other particles acquire their mass.

While it is beyond the scope of this paper to explain the mechanisms and exact details of how the Higgs mechanism works, it is integral to study and observe all of the properties in order to better understand the nature of mass and the interaction of fundamental particles. Studying and observations are accomplished through analyzing many events in supercolliders in which the Higgs boson has been produced. Therefore, possessing an accurate model that can label events as signal or background events is essential in learning more about the Higgs boson.

The Higgs boson has many different *channels* through which it can decay[9]. A channel is a way in which a particle may decay into certain specific final state particles. This decay produces other unique particles as a result of the instability of the Higgs boson. Three channels of Higgs decay have been observed, each consist of a decay to boson pairs. These discoveries only account for roughly ~40% of Higgs decays. The remaining ~60% of Higgs decay modes lies in the Higgs to bottom quark decay channel, shown below[9].

$$gg \rightarrow H^0 \rightarrow W^{\mp}H^{\pm} \rightarrow W^{\mp}W^{\pm}h^0 \rightarrow W^{\mp}W^{\pm}bb^- \qquad (1)$$

In this process, two gluons fuse into a Higgs boson. This boson then decays into a charged Higgs boson and a W boson. These particles then decay into two W bosons and a light Higgs boson. The light Higgs then decays into a pair of bottom and anti-bottom quark. The observance of these decay modes are necessary steps to confirming the generation of mass for fermions. The observations of the $H \rightarrow b\underline{b}$, are the decay modes modeled in the training data used by the classifier.

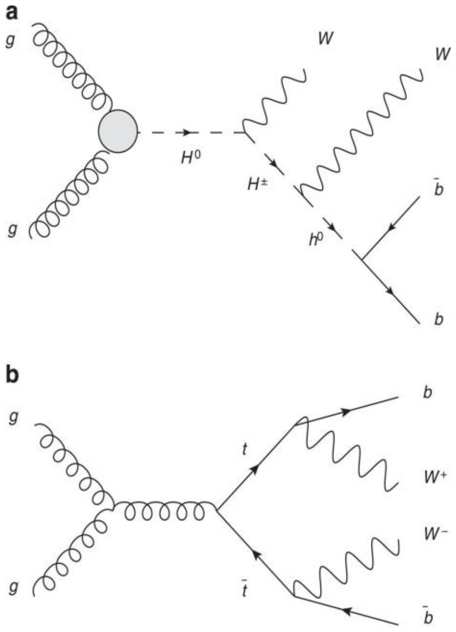

Figure 1a. $H \rightarrow b\underline{b}$ decay, Figure 1b. Background event

Figure 1. (a) Diagram describing signal event containing H^0 and $H^{\pm}$ bosons with decay congruent with $H \rightarrow b\underline{b}$.(b) Diagram describing a background event involving t-quarks, with same resulting $b\underline{b}$ decay with $w\underline{w}$.

The Large Hadron Collider provides the means of detecting the Higgs boson due to its capability to run searches for Higgs particles up to $1\ TeV/C^2$ with reactions of the type

$$p + p \rightarrow H^0 + X \qquad (2)$$

where X is any decay allowed by the conservation of energy laws. The mass of most Higgs bosons recently observed appears to be at $125\ GeV/c^2$[10]. The LHC circles proton-proton collisions at a rate of ~50 nanoseconds, these collisions are called events. When these protons collide, a shower of new particles is produced. The resulting particles can optionally decay into more stable particles or remain if stable[3]. One drawback to definitively proving the existence of certain particles is due to the unobservability of certain unstable intermediary particles. Physicists can only analyze the stable particles which hit the silicon layered detectors in the ATLAS and CMS detectors[11]. These detectors then use the evidence of the resulting particles to infer which intermediary particles were produced during specific events. Observable decay products that may be observed include electronically-charged leptons (ι), and particle jets.

The current approach to the identification of signal events is primarily through the direction of the resulting particles and their momentum[11]. The trigger system for these detectors automatically discards any unnecessary data; subsequently, it then makes the decision to keep certain events based on a three-layered classifier which significantly decreases the number of events written to disk, specifically from ~20,000,000 to ~400 events. The primary task is to aid in deciphering these ~400 events to determine the existence of a $H \ \rightarrow b\underline{b}$ event.

2.2 Machine Learning

Current standard approaches to the identification of Higgs bosons come through the standardized machine learning techniques[12]. Such techniques involve boosted decision trees and single-layer feed-forward neural networks. The shallow nature of these machine learning techniques requires significant effort to be put into manual feature extraction. These requirements significantly slow down and increase effort in the classification process.

New approaches have been proposed involving the use of deep neural networks in classifying event data[13]. This technique offers an improvement of 14% to 0.802 or 80% accuracy in classification. This study employed the use of a deep neural network with 300 hidden units in each of the five hidden layers and an initial learning rate of 0.03. This process marked a significant boost in the classification abilities of deep learning techniques.

Another approach was recently undertaken in using convolutional neural networks in identifying Higgs bosons[7]. This technique involved preprocessing significant amounts of data into jet energy images. This conversion allowed the use of computer vision tools to be utilized. Convolutional neural networks were used as classifiers trained off these jet images. This process raised the accuracy to 0.825 or 82%, a significant improvement over other methods. However, the preprocessing nature of the data and conversion to jet-pull images can be relatively costly and time-consuming compared to training with raw collider data.

2.3 Semi-Supervised Learning

Semi-supervised learning seeks to combine the efficient classification abilities of supervised learning with the self-sufficient labeling abilities of unsupervised learning. Fully labeled datasets are expensive and sometimes time-consuming to fully gather both data and labels. Because of this, an efficient way of training an accurate model can be achieved through semi-supervised learning. In our case, almost all data used comes with a label, but the abilities of semi-supervised learning techniques may yield a more accurate classification than traditional methods. Therefore, a portion of the data used is purposely de-labeled to exploit the semi-supervised learning methods, in addition to showcasing the application of SSL in high energy physics.

Recent work called MixMatch[14] has been published which proposes the use of a holistic approach to SSL in combining multiple learning techniques from past research. This method works by adding a loss term computed from the unlabeled data which encourages the model to better generalize to unlabeled data. The loss term is the specific term which is unified from three main areas: generic regularization- whose goal is to generalize and avoid the neural network from overfitting[15], consistency regularization - whose goal is to encourage confident output from the neural network when the input is transformed[16], and entropy minimization - whose goal is to encourage confident output from unlabeled data[17].

In order to properly disseminate the underlying methods used in MixMatch, it will be important to outline existing methods of SSL from which MixMatch builds upon.

2.3.1 consistency regularization

This technique relies heavily on data augmentation, the application of many different stochastic transformations on the input data, largely known to not have any effect on the important semantics of the data. Consistency regularization utilizes the effectiveness of data augmentation by exploiting the theorem that the model should output the same y over an x which has gone through augmentation Augment(x)[16].

$$||P_{model}(y|Augment(x);\ \theta)\ -\ P_{model}(y|Augment(x);\ \theta)||^{2}_{2} \tag{3}$$

It should be noted that Augment(x) is a stochastic transformation to the data, so each term is different[16]. One main use of the consistency regularization is from the application of the “Mean Teacher”[16] where one of the Augmentation terms is replaced with an exponential moving average (EMA) of the model’s parameters. The use of EMA enables a more accurate model than using the final parameters directly[18].

2.3.2 exponential moving average

The exponential moving average, classically used for time-series data, is used on the weights (parameters) of the neural network. The EMA is defined as:

$$\theta'_t \; = \; \alpha\, \theta'_{t-1} + (1-\alpha)\, \theta_t \tag{4}$$

With θ'_t at time t as a mix between the value of the raw weight θ_t mixed with the previous moving average θ'_{t-1}. This degree of mixing between the previous EMA and the current weight is determined by the smoothing coefficient hyperparameter α. This smooths the weights from noisy data, and gives the model a better prediction due to the smoothing.

2.3.3 entropy minimization

It is well known that a model's output decision boundary (the line between what is and is not a signal event) should not cross through areas of high-density in the data distribution[16]. This is achieved through the requirement that the model should output low-entropy y' on data points without labels. MixMatch's "Pseudo-Label"[14] method does this by producing 1-hot labels from high-confidence predictions on unlabeled data points. These 1-hot labels are then used as training targets employing cross-entropy loss functions.

2.3.4 general regularizatio

Regularization in machine learning largely refers to the effort of constraining the data and model in order to increase difficulty in the memorization of training data. This lets the model focus on the more abstract and important aspects of the data. This is largely done through weight decay, a process that penalizes the Euclidean norm of the model's parameters[19].

The power of the methods proposed in MixMatch is through the amalgamation of various effective semi-supervised learning methods into one holistic approach. It is for these reasons that MixMatch was chosen as the main implementation method from which a computational model may be developed that is effective in accurate classification $H{-}{>}\ b\underline{b}$ tagging with limited amounts of labeled data.

3 Data and Features

The Large Hadron collides protons from ionized helium every 50 nanoseconds within each of its four experiments[20]. These proton-proton collisions are called events. Within each event, the two colliding protons collide so rapidly that the quarks and gluons which make up the proton interact with enough energy to produce excitations in other fields, thus producing massive fundamental particles from the collision[3]. Because nature tends towards the lowest energy state, most of these massive intermediary particles are unstable and decay rapidly into less-energy final state particles. Multiple layers of detectors surround the point of collision in order to analyze different kinematic properties of these final-state particles. These kinematic data-points are the low-level features which are then used to train the computational model to differentiate between an event that is known to have produced a Higgs boson, and an event that has identical final-state particles but did not produce the Higgs boson.

Observable decay products of a proton-proton collision include charged leptons and particle jets. Particle jets are focused cones of particles originating from the hadronization of quarks or gluons[3]. The particle jets are extremely important because the signatures let us analyze many different properties, such as the momentum of the initial particle from where the jet originated. We analyze the momentum via three measurements: the momentum transverse to the beam direction, the polar angle, and the azimuthal angle. In addition to these three, we also use the pseudorapidity, a normalized version of the polar angle, as well as the momentum carried by the intermediate particle. While this may not be directly measured by the detectors, it can be inferred from the plane transverse to momentum using conservation of momentum. The residual imbalance of the transverse momentum must be due to the production of the intermediate particles.

With perfect measurements of the various kinematic properties of the final state particles B and C, we can calculate the invariant mass of the short-lived intermediate state A in process $A \rightarrow B + C$ with a relativistic 4-vector kinematic formula:

$$m_A{}^2 = m_{B+C}{}^2 = (E_B + E_C)^2 - |(p_B + p_C)|^2 \tag{5}$$

Where E is the energy and p is the momentum. Despite this simple formula, other factors such as escaping neutrinos, make it impossible to derive the precise mass of the intermediate properties. Instead, solely the kinematic data from the final state particles are studied via machine learning packages and are then used to derive the intermediate mass from such low-level data[21].

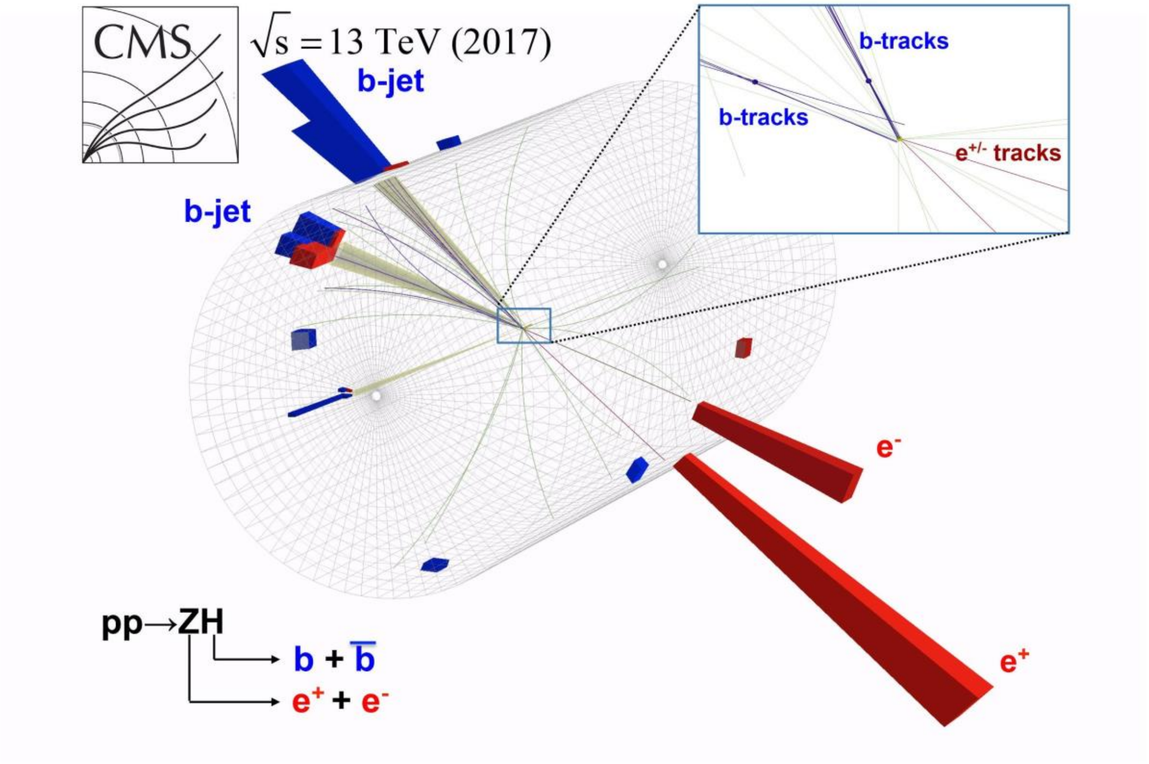

Figure 2. Signal event

Figure 2. Signal event captured by the CMS detector at the Large Hadron Collider. Shows the proper jet decay congruent with a Higgs boson through the $pp \rightarrow ZH$.

The data used to train the computational semi-supervised model in our research is produced via simulated events from the official CMS full detector simulator[22]. The simulator runs random proton-proton collisions based on all the knowledge and data accumulated at CERN. The simulator then produces kinematic data and virtual intermediary and final state particles. The resultant final state particles are tracked through the virtual model of the detector. Data is gathered from the kinematic properties measured of the final state particles. This process yields full simulated events with properties that mimic the statistical mechanics of real events.

4 Methods

4.1 Current Approach

There exist many standard machine learning methods and techniques available to physicists, such as the TMVA[12] package widely used by the physics community. Many of these machine learning methods rely on single-layer feed-forward neural networks and boosted decision trees. However, current research[10] shows that deep neural networks (more than one layer) trained on previous events produce higher accuracy ratings than those produced using the TMVA package. The single problem with these deep neural networks is the amount of labeled data needed to effectively train them. Therefore, we use state-of-the-art semi-supervised learning methods to attain similar accuracy ratings with just 50% labeled data.

4.2 Dataset

The data sets used as training, validation, and testing sets were derived from millions of particle collision simulations run by the CMS experiment. This data contains 3,640,000 total events, evenly balanced between signal and background events. Each event contains 172 features from which to identify whether the event produced a Higgs boson. From these 172 characteristics, we used only 27 features which we consider low-level features. These 27 characteristics make up most of the low-level kinematic data in order for the model to train without high-level features, obviating the need for complex equations in accurate identification.

The total training data points per training phase of the SSL computational model totals 182,000, of which only 50% to 0.5% were labeled. These were then split into 2843 mini-batches of 64 events.

4.3 MixMatch

The primary methods and techniques used in regard to development of the semi-supervised model come from MixMatch, a holistic approach to semi-supervised learning method which incorporates ideas from the dominant paradigms of SSL. With two batches X, U modeling the labeled examples and unlabeled examples respectively, MixMatch produces X' and U'. X' represents augmented labeled examples, while U' represents the unlabeled examples with their corresponding "guessed" label from which to train on.

For each unlabeled term in U, MixMatch produces a "guess" from the model's prediction. In order to use these labels, MixMatch computes the average of the models predicted class distribution across K augmentations using[14]:

$$Q_b = \frac{1}{K}\sum_{k=1}^{K} P_{model}(\, y \mid u_{b,k} \; ; \; \theta) \tag{6}$$

This artificial target is derived through this method. It should be noted that due to the nature of particle physics data for which we are applying MixMatch, augmentations are limited due to the fragility of the data used to train and test the model. Augmentations, as regarded in the original MixMatch paper, were designed for images, such as warping, rotating, etc. This spurred an effort to develop different techniques for augmentations. Most augmentations to the collider data involve adding differing amounts of noise, such as Gaussian distribution noise. Gaussian noise was applied with a mean of 0 and a standard deviation of 1. In addition to adding various amounts of noise to the input data, we also added noise to the activations of neurons, parameters of the model, and outputs. This addition increased robustness of the model in conjunction with reducing the possibility of overfitting the model.

A sharpening function is added as an additional step to the production of a 1-hot label from an unlabeled event. This function uses the approach suggested in *Miyato et al*[23] where we adjust the "temperature" of this categorical distribution:[14]

$$Sharpen(p,T)_i := p_i^{\frac{1}{T}} / \sum_{j=1}^{L} p_j^{\frac{1}{T}} \tag{7}$$

Where p is the average class distribution over K augmentations, and T is a temperature hyperparameter.

4.3.1 loss function

In order to adhere with the methods proposed with MixMatch, the loss terms were defined and used as[14]:

$$L_X = \frac{1}{|X'|} \sum_{p \in x'}^{x} H(p, P_{model}(y \mid x ; \theta) \tag{8}$$

$$L_U = \frac{1}{L|U'|} \sum_{p \in U'}^{u} || \, q - P_{model}(y \mid y ; \theta) ||_2^{\,2} \tag{9}$$

$$L = L_X + \lambda_U L_U \tag{10}$$

H is the standard PyTorch implementation cross-entropy loss function between target distribution p and q with the rest being hyperparameters.

The choice to use MixMatch as our semi-supervised learning techniques comes from the idea that its methods combine the most effective and state-of-the-art SSL techniques into one implementation.

4.4 Hyperparameter Optimization

Hyperparameters were chosen from different combinations of the parameters listed in Table 1. The combinations were chosen based on previous classification accuracy measured through the validation data during the training of the deep neural networks. The most accurate models based on testing data were the models with the largest amounts of layers and neurons per layer.

Table 1. Hyperparameters for computational models

Hyperparameters	**Options**
Number of Layers	3, 4, 5, 6
Hidden units per layer	128, 256, 512
Learning Rate	0.03, 0.01, 0.003, 0.001
Dropout	0.1, 0.3
Weight Decay	0.01, 0.001, 0.003

Other neural network hyperparameters were not determined through optimization methods. The ReLU[24] activation function was used for all hidden units, while a SoftMax activation function[25] was used for the output layer.

4.5 Computation

Computations were performed using machines with 8 Intel Xeon cores, an integrated Intel graphics processor, and 16 GB memory. All gradient computations were made on mini batches from 64 to 1024. All neural networks were trained using the PyTorch machine learning framework[26].

5 Results

We evaluate the performance of the semi-supervised learning computational models in terms of error rate and accuracy with varying number of labeled examples from 1000 to a completely labeled set. Each model was tested on data not seen in the training phase. The evaluation of neural networks with different numbers of hidden layers and hidden units with various amounts of labeled data points can also be seen in Figure 3. The exact accuracy scores can be viewed in Table 2.

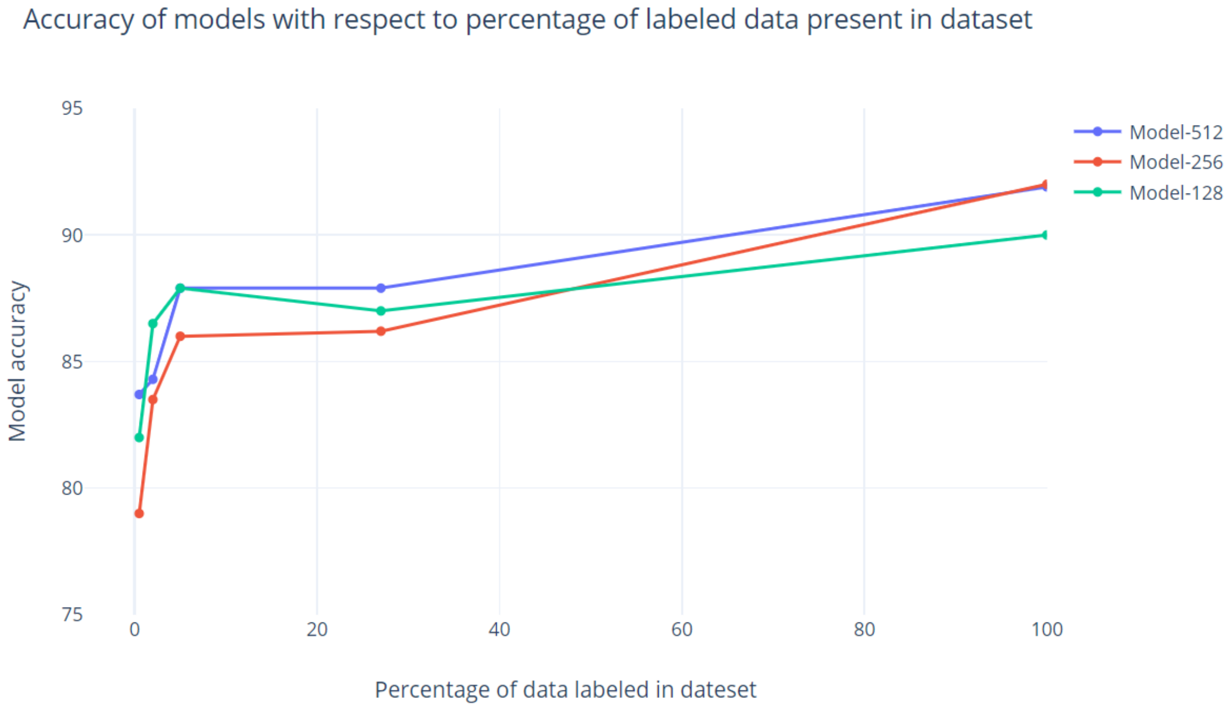

Figure 3. Accuracy of models with respect to amount of labeled data present in dataset

Figure 3. Evaluation of neural networks with different amounts of hidden layers and units. Models were tested on 0.5, 2, 5, 27, 100 percent labeled data in the dataset. As shown, the model with 512 units per hidden layer performed the best on 1000 and 50,000 labeled data points. Curiously, the model with 128 hidden units vastly outperformed other models on the dataset where only 5000 data points were labeled.

In total, the different models outperformed expectations on training with less than 1% of the data being labeled. Achieving close to an average 83% accuracy from the best model. According to results it seems that with any part of the data abstracted from the set results in the model having difficulty in surpassing 88% percent accuracy, except for a fully labeled dataset.

Table 2. Table of all accuracy scores of different models

Percentage of data labeled	Model – 128 hidden units per layer	Model – 256 hidden units per layer	Model – 512 hidden units per layer
100%	90.09	92.15	91.97
27%	86.89	86.21	87.98
5%	87.79	86.02	87.74
2%	85.90	83.50	84.33
0.5%	82.78	79.00	83.73

Table 2. Accuracy of models

All models followed similar trends of error rate dropping as more labels are added to the dataset used in the training phase. It should be noted that while error rates did drop, the rates for 0.5% labeled data point training remained competitively accurate despite the absence of more labeled data.

In comparison with previously outlined work, the model performance with 100% labeled data performed at a higher level by 7 – 9.65 % accuracy compared to models detailed in section 2.2. When labeled data was restricted to only 2% of the full dataset, the models proposed in this paper achieved similar accuracy ratings to previous work. Thus, demonstrating the ability of semi-supervised learning algorithms deprived of fully labeled datasets in achieving similar accuracy scores to fully supervised models with complete labeled datasets.

6 Conclusion

Limitations in machine learning algorithms currently being used in high-energy physics lead to manually extracted high-level functions derived from low-level data, in addition to lower accuracy scores. This research mitigates these shortcomings by implementing deep semi-supervised neural networks in the high-level feature extraction process. It is demonstrated how we have mitigated the need for complete labeled datasets in training such models in addition to proposing novel applications of semi-supervised learning methods and techniques. The authors hope this will aid in physicists' abilities to accurately identify the Higgs boson, in addition to helping reduce the amount of labeled data used by the European Centre for Nuclear Research. We hope the results of this research stimulate the development of even more powerful semi-supervised learning classification methods, specifically in its application to high-energy physics, of which there remains much scope. This appears to be the first implementation of cutting edge semi-supervised learning methods in this specific application to high energy physics.

Future work includes the development of domain-specific semi-supervised learning algorithms designed to work solely with particle accelerators. In addition to this, future work would also include the application of semi-supervised learning towards individual particle tracking and event reconstruction.

7 Acknowledgments

We are grateful for Dr. Sanft, Dr. Cameron, and the entire computer science department at the University of North Carolina at Asheville. We would like to thank the entire graduating class of Computer Science majors for their help and encouragement. In addition, we would like to thank Derek Kingma for his guidance in the area of Semi-Supervised Learning near the start of this research.

8 References

1. Albertsson, Kim, Piero Altoe, Dustin Anderson, John Anderson, Michael Andrews, Juan Pedro Araque Espinosa, Adam Aurisano et al. "Machine learning in high energy physics community white paper." *arXiv preprint arXiv:1807.02876* (2018).
2. Bertasius, Gedas, Jianbo Shi, and Lorenzo Torresani. "High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 504-512. 2015.
3. Thomson, Mark. *Modern particle physics*. Cambridge University Press, 2013.
4. Baldi, Pierre, Peter Sadowski, and Daniel Whiteson. "Searching for exotic particles in high-energy physics with deep learning." *Nature communications* 5 (2014): 4308.
5. Zhu, Xiaojin, and Andrew B. Goldberg. "Introduction to semi-supervised learning (synthesis lectures on artificial intelligence and machine learning)." *Morgan and Claypool Publishers* 14 (2009).
6. Sadowski, Peter J., Daniel Whiteson, and Pierre Baldi. "Searching for higgs boson decay modes with deep learning." In *Advances in Neural Information Processing Systems*, pp. 2393-2401. 2014.
7. Apostolatos, Anton, and Leonard Bronner. "Identifying the Higgs Boson with Convolutional Neural Networks."
8. Lyth, David H., and Ewan D. Stewart. "Cosmology with a TeV mass Higgs field breaking the grand-unified-theory gauge symmetry." *Physical Review Letters* 75, no. 2 (1995): 201.
9. ATLAS, Collaboration, Marco Agustoni, Hans Peter Beck, Alberto Cervelli, Antonio Ereditato, Sigve Haug, Sonja Kabana et al. "Search for the bb decay of the Standard Model Higgs boson in associated (W/Z) H production with the ATLAS detector." *Journal of High Energy Physics* 1, no. 1 (2015): 69.
10. Bezrukov, Fedor, Mikhail Yu Kalmykov, Bernd A. Kniehl, and Mikhail Shaposhnikov. "Higgs boson mass and new physics." *Journal of High Energy Physics* 2012, no. 10 (2012): 140..
11. Sirunyan, Albert M., and CMS collaboration. "Particle-flow reconstruction and global event description with the CMS detector." *JINST* 12, no. 10 (2017): P10003.
12. Hoecker, Andreas, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, Helge Voss, M. Backes et al. "TMVA-toolkit for multivariate data analysis." *arXiv preprint physics/0703039* (2007).
13. Sadowski, Peter J., Daniel Whiteson, and Pierre Baldi. "Searching for higgs boson decay modes with deep learning." In *Advances in Neural Information Processing Systems*, pp. 2393-2401. 2014.
14. Berthelot, David, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. "Mixmatch: A holistic approach to semi-supervised learning." In *Advances in Neural Information Processing Systems*, pp. 5050-5060. 2019.
15. Alpaydin, Ethem. *Introduction to machine learning*. MIT press, 2020.
16. Tarvainen, Antti, and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results." In *Advances in neural information processing systems*, pp. 1195-1204. 2017.
17. Grandvalet, Yves, and Yoshua Bengio. "Semi-supervised learning by entropy minimization." In *Advances in neural information processing systems*, pp. 529-536. 2005.
18. Papernot, Nicolas, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. "Distillation as a defense to adversarial perturbations against deep neural networks." In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582-597. IEEE, 2016.
19. Wager, Stefan, Sida Wang, and Percy S. Liang. "Dropout training as adaptive regularization." In *Advances in neural information processing systems*, pp. 351-359. 2013.
20. Bunn, Julian J., Harvey Newman, Shawn McKee, David G. Foster, Richard Cavanaugh, and Richard Hughes-Jones. "High speed data gathering, distribution and analysis for physics discoveries at the large Hadron collider." In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, pp. 241-es. 2006.
21. Collaboration, C. M. S. "The CMS experiment at the CERN LHC." (2008).
22. Duarte, Javier. “Sample with jet, track and secondary vertex properties for Hbb tagging ML studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC.” *CERN Open Data Portal.* DOI:10.7483/OPENDATA.CMS.JGJX.MS7Q. 2019

23. Miyato, Takeru, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. "Virtual adversarial training: a regularization method for supervised and semi-supervised learning." *IEEE transactions on pattern analysis and machine intelligence* 41, no. 8 (2018): 1979-1993.

24. Ramachandran, Prajit, Barret Zoph, and Quoc V. Le. "Searching for activation functions." *arXiv preprint arXiv:1710.05941* (2017).

25. Boutilier, Craig, Relu Patrascu, Pascal Poupart, and Dale Schuurmans. "Constraint-based optimization and utility elicitation using the minimax decision criterion." *Artificial Intelligence* 170, no. 8-9 (2006): 686-713.

26. Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen et al. "PyTorch: An imperative style, high-performance deep learning library." In *Advances in Neural Information Processing Systems*, pp. 8024-8035. 2019.