Deep Learning in High-Energy Physics: Improving the Search for Exotic Particles

P. Baldi,[1] P. Sadowski,[1] and D. Whiteson[2]

[1]*Dept. of Computer Science, UC Irvine, Irvine, CA 92617*
[2]*Dept. of Physics and Astronomy, UC Irvine, Irvine, CA 92617*

Collisions at high-energy particle colliders are a traditionally fruitful source of exotic particle discoveries. Finding these rare exotic particles requires solving difficult signal-versus-background classification problems, hence machine learning approaches are often used for this task. Standard approaches in the past have relied on 'shallow' machine learning models that have a limited capacity to learn complex non-linear functions of the inputs, and rely on a pain-staking search through manually constructed non-linear inputs. Progress on this problem has slowed, as a variety of techniques (neural networks, boosted decision trees, support vector machines) have shown equivalent performance. Recent advances in the field of deep learning, particularly with artificial neural networks, make it possible to learn more complex functions and better discriminate between signal and background classes. Using benchmark datasets, we show that deep learning methods need no manually constructed inputs and yet improve the AUC (Area Under the ROC Curve) classification metric by as much as 8% over the best current approaches. This is a large relative improvement and demonstrates that deep learning approaches can improve the power of collider searches for exotic particles.

The field of *high energy physics* is devoted to the study of the elementary constituents of matter. By investigating the structure of matter and the laws that govern its interactions, this field strives to discover the fundamental properties of the physical universe. The primary tools of experimental high energy physicists are modern accelerators, which collide protons and/or anti-protons to create exotic particles that occur only at extremely high energy densities. Observing these particles and measuring their properties may yield critical insights about the very nature of matter [1]. Such discoveries require powerful statistical methods, and machine learning tools play a critical role. Given the limited quantity and expensive nature of the data, improvements in analytical tools directly boost particle discovery potential.

To discover a new particle, physicists must isolate a subspace of their high-dimensional data in which the hypothesis of a new particle or force gives a significantly different prediction than the null hypothesis, allowing for an effective statistical test. For this reason, the critical element of the search for new particles and forces in high-energy physics is the computation of the *relative likelihood* of data under different hypotheses. Often this relative likelihood function cannot be expressed analytically, so simulated collision data generated with Monte Carlo methods are used as a basis for approximation of the likelihood function. The high dimensionality of data, referred to as the *feature space*, makes it intractable to generate enough simulated collisions to describe the relative likelihood in the full feature space, and machine learning tools are used for dimensionality reduction. Machine learning classifiers such as neural networks provide a powerful way to solve this learning problem, essentially by estimating the probability of membership in each class.

In this paper, we show that the current techniques used in high-energy physics fail to capture all of the available information, even when boosted by manually-constructed physics-inspired features. This effectively reduces the power of the collider to discover new particles. We demonstrate that recent developments in deep learning tools can overcome these failings, providing significant boosts even without manual assistance.

The relative likelihood function is a complicated function in a high-dimensional space. While any function can theoretically be represented by a 'shallow' classifier, such as a neural network with a single hidden layer [2], an intractable number of hidden units may be required. Circuit complexity theory tells us that deep neural networks have the potential to compute complex functions much more efficiently (fewer hidden units), but in practice they are notoriously difficult to train due to the vanishing gradient problem [3, 4]; the adjustments to the weights in the early layers of a deep network rapidly approach zero during training. A common approach is to combine shallow classifiers with high-level features that are derived manually from the raw features. These are generally non-linear functions of the input features that capture physical insights about the data. While helpful, this approach is labor-intensive and not necessarily optimal; a robust machine learning method would obviate the need for this additional step and capture all of the available classification power directly from the raw data.

Recent successes in deep learning – *e.g.* neural networks with multiple hidden layers – have come from alleviating the gradient diffusion problem by a combination of factors, including: 1) speeding up the stochastic gradient descent algorithm with graphics processors; 2) using much larger training sets; 3) using new learning algorithms, including randomized algorithms such as dropout [5, 6]; and 4) pre-training the initial layers of the network with unsupervised learning methods such

as autoencoders [7, 8]. With these methods, it is becoming common to train deep networks of five or more layers. These advances in deep learning could have a significant impact on applications in high-energy physics. Construction and operation of the particle accelerators is extremely expensive, so any additional classification power extracted from the collision data is very valuable.

RESULTS

High Energy Physics Benchmarks

The vast majority of particle collisions do not produce particles of interest. For example, though the Large Hadron Collider produces approximately 10^{11} collisions per hour, approximately 300 of these collisions result in a Higgs boson, on average. Therefore, good data analysis depends on effective collision selection, in which events producing particles of interest (*signal*) are separated from those producing other particles (*background*).

Even when interesting particles are produced, detecting them poses considerable challenges. They are too small to be directly observed and decay almost immediately into other particles. Though new particles cannot be directly observed, the lighter stable particles to which they decay, called *decay products*, can be observed. Multiple layers of detectors surround the point of collision for this purpose. As each decay product pass through these detectors, it interacts with them in a way that allows its direction and momentum to be measured.

Observable decay products include electrically-charged leptons (electrons or muons, denoted ℓ), and particle jets (collimated streams of particles originating from quarks or gluons, denoted j). In the case of jets we attempt to distinguish between jets from heavy quarks (b) and jets from gluons or low-mass quarks; jets consistent with b-quarks receive a b-quark *tag*. For each object, the momentum is determined by three measurements: the momentum transverse to the beam direction (p_{T}), and two angles, η (polar) and ϕ (azimuthal). Finally, an important quantity is the amount of momentum carried away by the invisible particles. This cannot be directly measured, but can be inferred in the plane transverse to the beam by requiring conservation of momentum. The initial state has zero momentum away from the beam axis, therefore any imbalance of transverse momentum (denoted $\not{E}_T$) in the final state must be due to production of invisible particles such as neutrinos (ν) or exotic particles.

Benchmark Case 1: Higgs Bosons (HIGGS)

The first benchmark classification task is to distinguish between a signal process where new theoretical Higgs bosons are produced, and a background process with the identical decay products but distinct kinematic features. This benchmark task was recently considered by experiments at the LHC [9] and the Tevatron colliders [10].

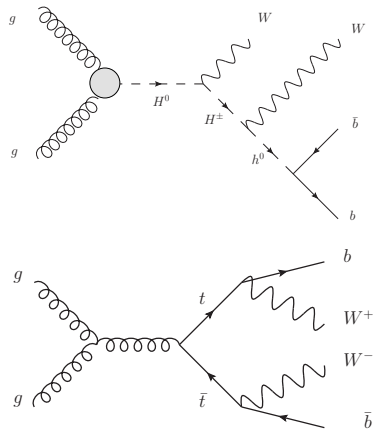

FIG. 1: Diagrams describing the signal process involving new exotic Higgs bosons ($H^0, H^{\pm}$, left) and the background process involving top-quarks (t, right). In both cases, the resulting particles are two W bosons and two b-quarks.

The signal process is the fusion of two gluons into a heavy electrically-neutral Higgs boson ($gg \to H^0$), which decays to a heavy electrically-charged Higgs bosons ($H^{\pm}$) and a W boson. The $H^{\pm}$ boson subsequently decays to a second W boson and the light Higgs boson, h^0 which has recently been observed by the ATLAS [11] and CMS [12] experiments. The light Higgs boson decays predominantly to a pair of bottom quarks, giving the process:

$$gg \to H^0 \to W^{\mp}H^{\pm} \to W^{\mp}W^{\pm}h^0 \to W^{\mp}W^{\pm}b\bar{b},$$

which leads to $W^{\mp}W^{\pm}b\bar{b}$, see Figure 1. The background process, which mimics $W^{\mp}W^{\pm}b\bar{b}$ without the Higgs boson intermediate state, is the production of a pair of top quarks, each of which decay to Wb, also giving $W^{\mp}W^{\pm}b\bar{b}$, see Figure 1.

Simulated events are generated with the MADGRAPH5 [13] event generator, with showering and hadronization performed by PYTHIA [14] and detector response simulated by DELPHES [15]. For the benchmark case here, $m_{H^0} = 425$ GeV and $m_{H^{\pm}} = 325$ GeV.

We focus on the semi-leptonic decay mode, in which one W boson decays to a lepton and neutrino ($\ell\nu$) and the other decays to a pair of jets (jj), giving decay products $\ell\nu b\ jjb$. We consider events which satisfy the requirements:

- Exactly one electron or muon, with $p_T > 20$ GeV and $|\eta| < 2.5$;
- at least four jets, each with $p_T > 20$ GeV and $|\eta| < 2.5$;

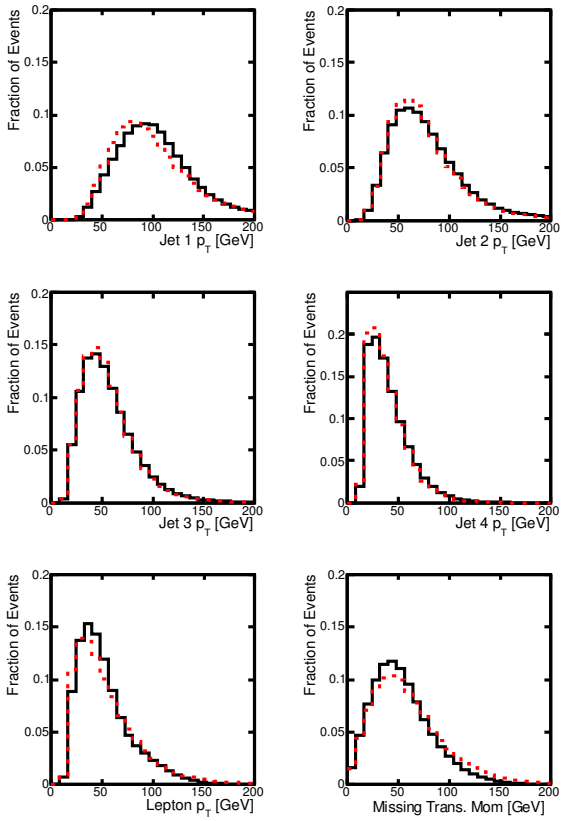

FIG. 2: Low-level input features from basic kinematic quantities in $\ell\nu jjb\bar{b}$ events for simulated signal (black) and background (red) benchmark events. Shown are the transverse momenta (p_{T}) of each observed particle as well as the imbalance of momentum in the final state. Momentum angular information for each observed particle is also available to the network, but is not shown, as the one-dimensional projections have little information.

- b-tags on at least two of the jets, indicating that they are likely due to b-quarks rather than gluons or lighter quarks.

Events which satisfy the requirements above are naturally described by a simple set of features which represent the basic measurements made by the particle detector: the momentum of each observed particle. In addition we reconstruct the missing transverse momentum in the event and have b-tagging information for each jet. Together, these twenty-one features comprise our *low-level feature set*. Figure 2 shows the distribution of a subset of these kinematic features for signal and background processes.

The low-level features show some distinguishing characteristics, but our knowledge of the different intermediate states of the two processes allows us to construct other features which better capture the differences. As the difference in the two hypotheses lies mostly in the existence of new intermediate Higgs boson states, we can distinguish between the two hypotheses by attempting to identify whether the intermediate state existed. This is done by reconstructing its characteristic invariant mass; if a particle A decays into particles B and C, the invariant mass of particle A (m_A) can be recovered as:

$$m_A^2 = m_{B+C}^2 = (E_B + E_C)^2 - |(\mathbf{p}_B + \mathbf{p}_C)|^2$$

where E is the energy and $\mathbf{p}$ is the three-dimensional momentum of the particle. In the signal hypothesis we expect that:

- $W \to \ell\nu$ gives a peak in $m_{\ell\nu}$ [16] at m_W,
- $W \to jj$ gives a peak in m_{jj} at m_W,
- $h^0 \to b\bar{b}$ gives a peak in $m_{b\bar{b}}$ at m_{h^0},
- $H^\pm \to Wh^0$ gives a peak in $m_{Wb\bar{b}}$ at $m_{H^\pm}$,
- $H^0 \to WH^\pm$ gives a peak in $m_{WWb\bar{b}}$ at m_{H^0},

while in the case of the $t\bar{t}$ background we expect that:

- $W \to \ell\nu$ gives a peak in $m_{\ell\nu}$ at m_W,
- $W \to jj$ gives a peak in m_{jj} at m_W,
- $t \to Wb$ gives a peak in $m_{j\ell\nu}$ and m_{jbb} at m_t.

See Figure 3 for distributions of these high-level features for both signal and background processes. Clearly these contain more discrimination power than the low-level features.

We have published a dataset containing 11 million simulated collision events for benchmarking machine learning classification algorithms on this task, which can be found in the UCI Machine Learning Repository [17] at archive.ics.uci.edu/ml/datasets/HIGGS.

Benchmark Case 2: Supersymmetry Particles (SUSY)

The second benchmark classification task is to distinguish between a process where new supersymmetric particles are produced, leading to a final state in which some particles are detectable and others are invisible to the experimental apparatus, and a background process with the same detectable particles but fewer invisible particles and distinct kinematic features. This benchmark problem is currently of great interest to the field of high-energy physics, and there is a vigorous effort in the literature [18–21] to build high-level features which can aid in the classification task.

The signal process is the production of electrically-charged supersymmetric particles ($\chi^\pm$), which decay to W bosons and an electrically-neutral supersymmetric particle χ^0, which is invisible to the detector. The W

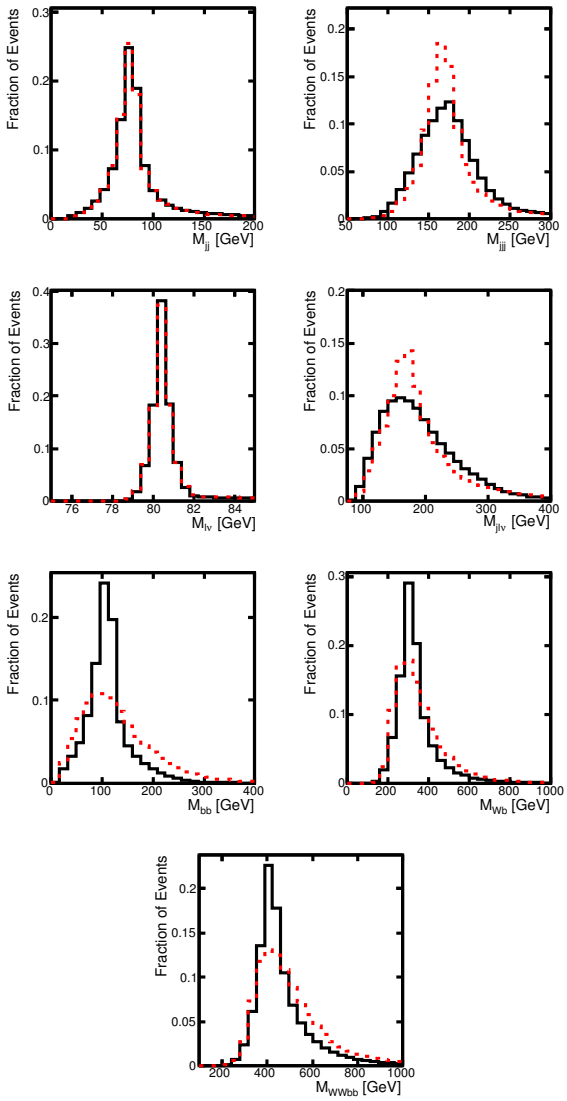

FIG. 3: High-level input features from invariant mass calculations in $\ell\nu jjb\bar{b}$ events for simulated signal (black) and background (red) events.

bosons decay to charged leptons ℓ and invisible neutrinos ν, see Fig. 4. The final state in the detector is therefore two charged leptons ($\ell\ell$) and missing momentum carried off by the invisible particles ($\chi^0\chi^0\nu\nu$). The background process is the production of pairs of W bosons, which decay to charged leptons ℓ and invisible neutrinos ν, see Fig. 4. The visible portion of the signal and background final states both contain two leptons ($\ell\ell$) and large amounts of missing momentum due to the invisible particles. The classification task requires distinguishing between these two processes using the measurements of the charged lepton momenta and the missing transverse momentum.

As above, simulated events are generated with the MADGRAPH [13] event generator, with showering and

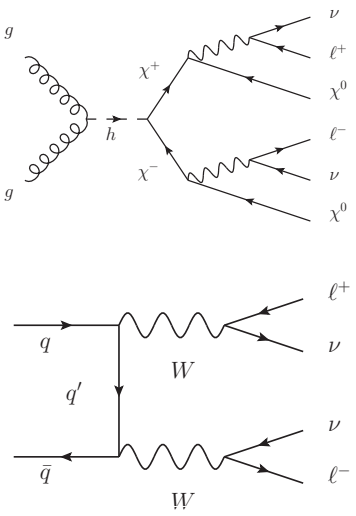

FIG. 4: Example diagrams describing the signal process involving hypothetical supersymmetric particles $\chi^{\pm}$ and χ^0 along with charged leptons $\ell^{\pm}$ and neutrinos ν (top) and the background process involving W bosons (bottom). In both cases, the resulting observed particles are two charged leptons, as neutrinos and χ^0 escape undetected.

hadronization performed by PYTHIA [14] and detector response simulated by DELPHES [15]. The masses are set to $m_{\chi^{\pm}} = 200$ GeV and $m_{\chi^0} = 100$ GeV.

We focus on the fully leptonic decay mode, in which both W bosons decay to charged leptons and neutrinos, $\ell\nu\ell\nu$. We consider events which satisfy the requirements:

- Exactly two electrons or muons, each with $p_T > 20$ GeV and $|\eta| < 2.5$;
- at least 20 GeV of missing transverse momentum

As above, the basic detector response is used to measure the momentum of each visible particle, in this case the charged leptons. In addition, there may be particle jets induced by radiative processes. A critical quantity is the missing transverse momentum, $\not{E}_T$. Figure 5 gives distributions of low-level features for signal and background processes.

Discovery of supersymmetry is a central piece of the scientific mission of the Large Hadron Collider. The strategy we applied to the Higgs boson benchmark, of reconstructing the invariant mass of the intermediate state, is not feasible here, as there is too much information carried away by the escaping neutrinos (two neutrinos in this case, compared to one for the Higgs case). Instead, a great deal of intellectual energy has been spent in attempting to devise features which give additional classification power. These include high-level features such as:

- Axial $\not{E}_T$: missing transverse energy along the vector defined by the charged leptons,

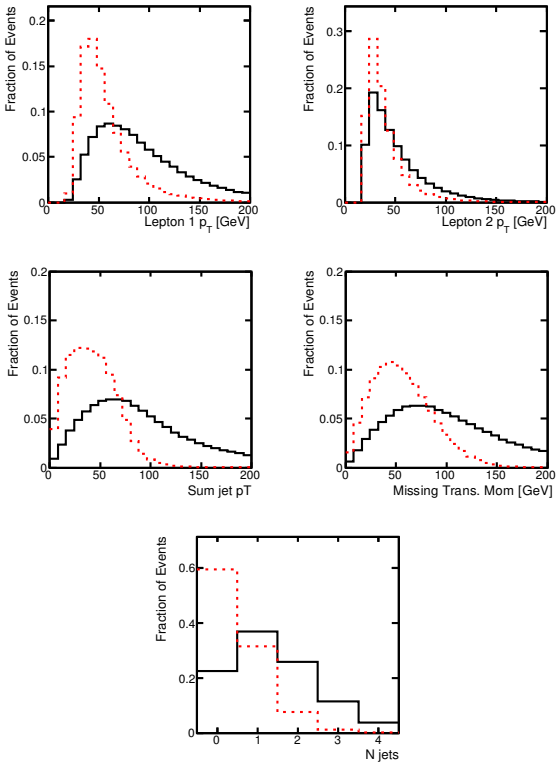

FIG. 5: Distribution of low-level features in simulated samples for the SUSY signal (black) and background (red) benchmark processes.

- stransverse mass M_{T2}: estimating the mass of particles produced in pairs and decaying semi-invisibly [18, 19],
- $\not{E}_T{}^{Rel}$: $\not{E}_T$ if $\Delta\phi \geq \pi/2$, $\not{E}_T \sin(\Delta\phi)$ if $\Delta\phi < \pi/2$, where $\Delta\phi$ is the minimum angle between $\not{E}_T$ and a jet or lepton,
- razor quantities β,R, and M_R [20],
- super-razor quantities β_{R+1}, $\cos(\theta_{R+1})$, $\Delta\phi_R^\beta$, M_Δ^R, M_R^T, and $\sqrt{\hat{s}_R}$ [21].

See Figure 6 for distributions of these high-level features for both signal and background processes.

A dataset containing five million simulated collision events are available for download at archive.ics.uci.edu/ml/datasets/SUSY.

Current Approach

Standard techniques in high-energy physics data analyses include feed-forward neural networks with a single hidden layer and boosted-decision trees. We use the

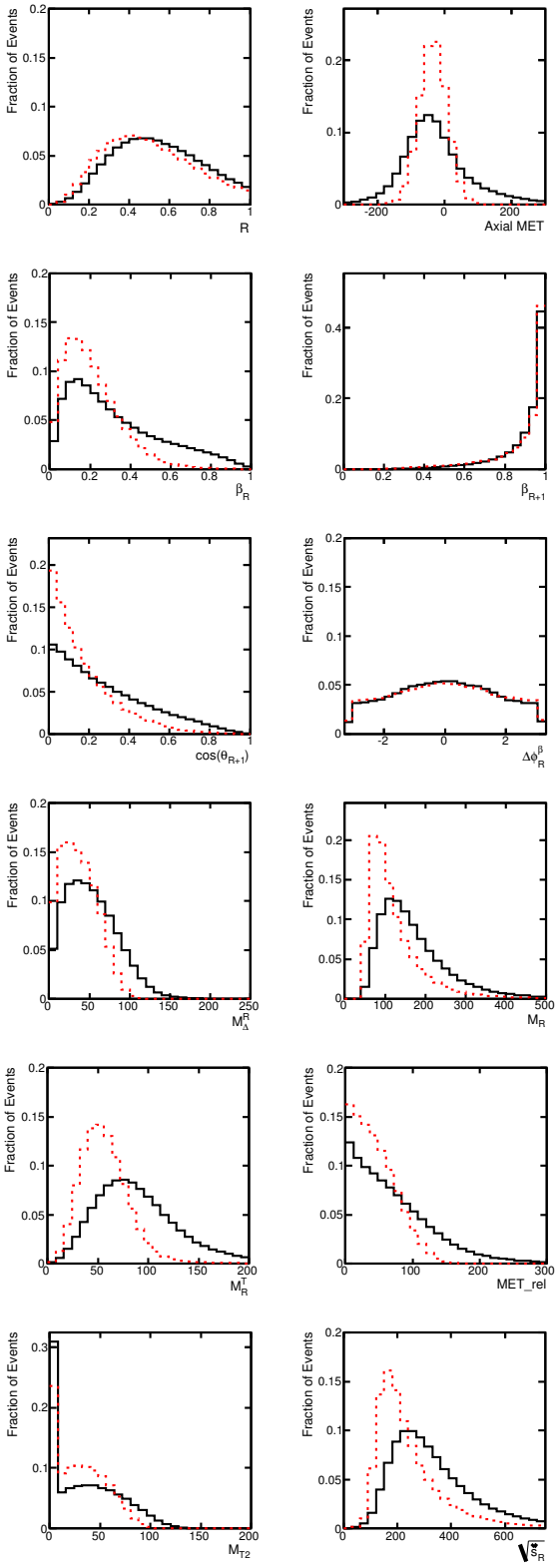

FIG. 6: Distribution of high-level features in simulated samples for the SUSY signal (black) and background (red) benchmark processes.

TABLE I: Hyper-parameter choices

Hyper parameters	Choices
Depth	2,3,4,5,6 layers
Hidden units per layer	100,200,300,500
Learning rate	0.01, 0.05
Weight decay	0, 0.00001
Pre-training	none, autoencoder
	multi-task autoencoder
Input features	low-level, high-level
	complete set

widely-used TMVA package [22], which provides a standardized implementation of common multi-variate learning techniques and an excellent performance baseline.

Deep Learning

Deep neural networks can be trained with a number of different algorithms and hyper-parameters. To make these choices under computational constraints, we performed a limited optimization on a small subset of the HIGGS data consisting of 2.6 million training examples and 100,000 validation examples, in which we compared the pre-training methods, training algorithms, and network architectures shown in Table I. The best parameter combination in our analysis was a five-layer neural network with the complete feature set as input, 300 hidden units in each hidden layer, a learning rate of 0.05, and a weight decay coefficient of 1×10^{-5}. We chose not to use pretraining, as it made no difference in the validation performance.

The chosen hyper-parameters were used to train both deep and shallow neural networks (with a single hidden layer) on both benchmark data sets. To investigate whether the neural networks were able to learn the discriminative information contained in the high-level features, we also trained separate classifiers for each of the three feature sets described above: low-level, high-level and combined features. Additional details are provided in the Methods section below.

Performance

Classifiers were tested on 500,000 simulated examples generated from the same Monte Carlo procedures as the training sets. We produced Receiver Operating Characteristic (ROC) curves to illustrate the performance of the classifiers. Our primary metric for comparison is the area under the ROC curve (AUC), with larger AUC values indicating higher classification accuracy across a range of threshold choices.

TABLE II: Comparison of the performance of several learning techniques for three sets of input features: low-level features, high-level features and the complete set of features. Each neural network was trained five times with different random initializations. The table displays the mean Area Under the Curve (AUC) of the signal-rejection curve in Figure 7, with standard deviations in parentheses.

	AUC		
Technique	Low-level	High-level	Complete
BDT	0.73	0.78	0.81
NN	0.733 (0.007)	0.777 (0.001)	0.816 (0.004)
DN	0.880 (0.001)	0.800 (< 0.001)	0.885 (0.002)

There are other choices of metrics one could consider, such as signal efficiency at some fixed background rejection, or discovery significance as calculated by p-value in the null hypothesis. We choose AUC as it is a standard in machine learning, and is closely correlated with the other metrics. Note, in addition, that small increases in AUC can represent significant enhancement in discovery significance.

Benchmark 1: HIGGS

Figure 7 and Table II show the signal efficiency and background rejection for varying thresholds on the output of the neural network (NN) or boosted decision tree (BDT).

A shallow NN or BDT trained using only the low-level features performs significantly worse than one trained with only the high-level features. This implies that the shallow NN and BDT are not succeeding in independently discovering the discriminating power of the high-level features. This is a well-known problem with shallow learning methods, and motivates the calculation of high-level features.

Methods trained with only the high-level features, however, have a weaker performance than those trained with the full suite of features, which suggests that despite the insight represented by the high-level features, they do not capture all of the information contained in the low-level features. The deep learning techniques show nearly equivalent performance using the low-level features and the complete features, suggesting that they are *automatically discovering the insight contained in the high-level features*. Finally, the deep learning technique finds additional separation power beyond what is contained in the high-level features, demonstrated by the superior performance of the deep network with low-level features to the traditional network using high-level features. These results demonstrate the advantage to using deep learning techniques for this type of problem.

The internal representation of a NN is notoriously dif-

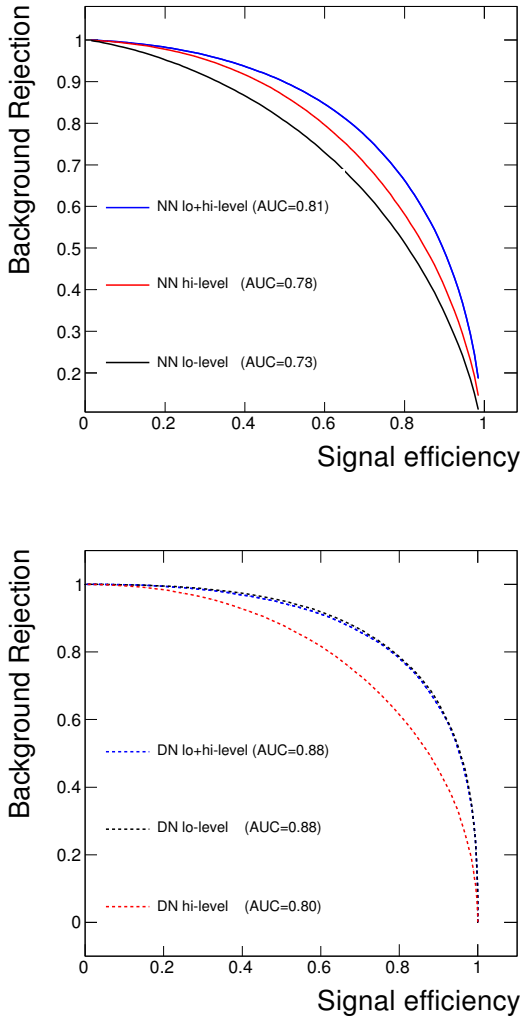

FIG. 7: Comparison of background rejection versus signal efficiency for the traditional learning method (top) and the deep learning method (bottom) using the low-level features, the high-level features and the complete set of features.

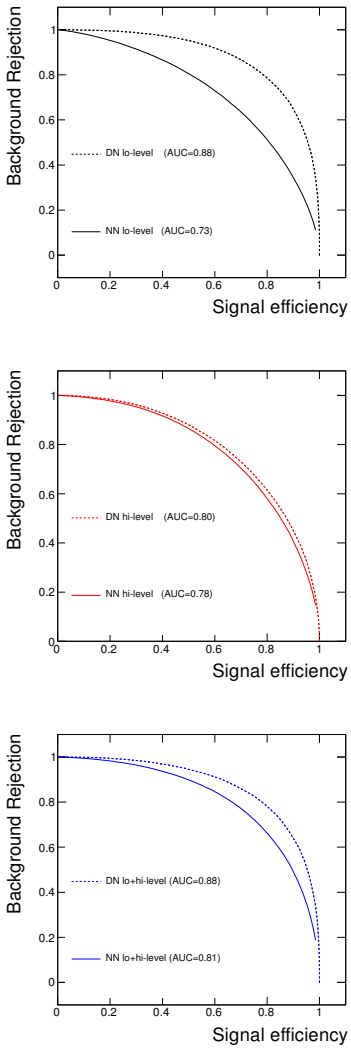

FIG. 8: Comparison of background rejection versus signal efficiency for the low-level features (top), high-level features (center) and complete set of features (bottom) using traditional and deep learning methods.

ficult to reverse engineer. To gain some insight into the mechanism by which the deep network (DN) is improving upon the discrimination in the high-level physics features, we compare the distribution of simulated events selected by a minimum threshold on the NN or DN output, chosen to give equivalent rejection of 90% of the background events. Figure 9 shows events selected by such thresholds in an even mixture of signal and background collisions, compared to pure distributions of signal and background. The NN preferentially selects events with values of the features close to the characteristic signal values and away from background-dominated values. The DN, which has a higher efficiency for the equivalent rejection, selects events near the same signal values, but *also retains events away from the signal-dominated region*. The likely explanation is that the DN has discovered the same signal-rich region identified by the physics features, but has in addition found avenues to carve into the background-dominated region.

Benchmark 2: Supersymmetry Particles (SUSY)

In the case of the SUSY benchmark, deep neural networks again perform better than shallow networks. The improvement is less dramatic, though statistically significant.

An additional boost in performance is obtained by using the dropout training algorithm, in which we stochastically drop neurons in the top hidden layer with 50%

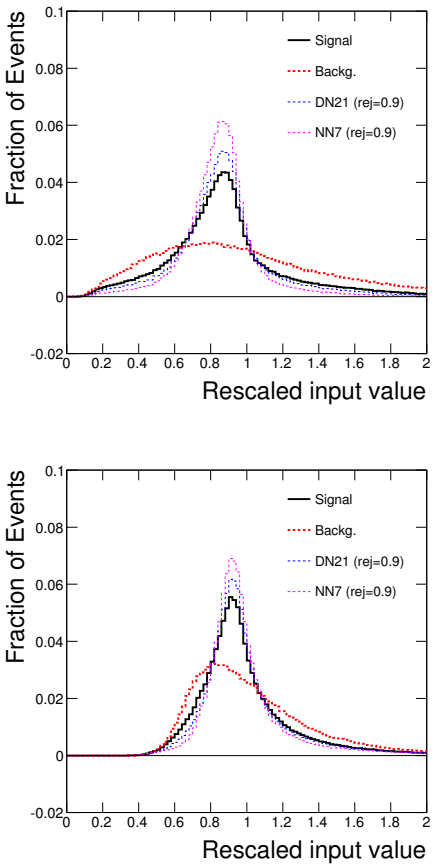

FIG. 9: Distribution of events for two rescaled input features: m_{Wbb} and m_{WWbb}. Shown are pure signal and background distributions, as well as events which pass a threshold requirement which gives a background rejection of 90% for a deep network with 21 low-level inputs (DN21) and a shallow network with 7 high-level inputs (NN7).

probability during training. For deep networks trained with dropout, we achieve an AUC of 0.88 on both the low-level and complete feature sets. Table III compares the results for different network depths, input features, and whether or not we use the dropout algorithm. The hyperparameters chosen for the Higgs benchmark were re-used for all these models.

In this SUSY case, neither the high-level features nor the deep network finds dramatic gains over the shallow network of low-level features. This suggests that the optimal separation function may be closer to a linear combination of the low-level features for this benchmark than for others, demonstrated by the performance of the shallow network. The power of the deep network to automatically find non-linear features reveals something about the nature of the classification problem in this case: the deep-network analysis argues that there may be little gain from further attempts to manually construct high-level features.

TABLE III: Performance comparison for the SUSY benchmark. Each model was trained five times with different weight initializations. The mean AUC is shown with the standard deviation in parentheses.

	AUC		
Technique	Low-level	High-level	Complete
BDT	0.850 (0.003)	0.835 (0.003)	0.863 (0.003)
NN	0.867 (0.002)	0.863 (0.001)	0.875 (< 0.001)
$\text{NN}_{dropout}$	0.856 (< 0.001)	0.859 (< 0.001)	0.873 (< 0.001)
DN	0.872 (0.001)	0.865 (0.001)	0.876 (< 0.001)
$\text{DN}_{dropout}$	0.876 (< 0.001)	0.869 (< 0.001)	0.879 (< 0.001)

DISCUSSION

It is widely accepted in experimental high-energy physics that machine learning techniques can provide powerful boosts to searches for exotic particles. Until now, physicists have reluctantly accepted the limitations of the shallow networks employed to date; in an attempt to circumvent these limitations, physicists manually construct helpful non-linear feature combinations to guide the shallow networks.

Our analysis shows that recent advances in deep learning techniques may lift these limitations by **automatically discovering optimal non-linear feature combinations and providing better discrimination power than current classifiers** – even when aided by manually-constructed features.

We suspect that the novel environment of high energy physics, with high volumes of relatively low-dimensional data containing rare signals hiding under enormous backgrounds, can inspire new developments in machine learning tools. Beyond these simple benchmarks, deep learning methods may be able to tackle thornier problems with multiple backgrounds, or lower-level tasks such as identifying the decay products from the high-dimensional raw detector output.

Acknowledgments

We are grateful to Kyle Cranmer, Chris Hays, Chase Shimmin, Davide Gerbaude, Bo Jayatilaka, Jahred Adelman, and Shimon Whiteson for their insightful comments. We wish to acknowledge a hardware grant from NVIDIA.

References

[1] Dawson, S. *et al.* Higgs Working Group Report of the Snowmass 2013 Community Planning Study (2013). [arXiv:1310.8361].

[2] Hornik, K., Stinchcombe, M. & White, H. *Neural Netw.* **2**, 359366 (1989).

[3] Hochreiter, S. *Recurrent Neural Net Learning and Vanishing Gradient* (1998).

[4] Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on* **5**, 157–166 (1994).

[5] Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors (2012). [arXiv:1207.0580].

[6] Baldi, P. & Sadowski, P. The dropout learning algorithm. *Artificial Intelligence* (2014). In press.

[7] Hinton, G. E., Osindero, S. & Teh, Y.-W. *Neural Computation* **18**, 1527–1554 (2006).

[8] Bengio, Y. *et al.* In *In NIPS* (MIT Press, 2007).

[9] Aad, G. *et al.* Search for a Multi-Higgs Boson Cascade in $W^+W b\bar{b}$ events with the ATLAS detector in pp collisions at s = 8 TeV (2013). [arXiv:1312.1956].

[10] Aaltonen, T. *et al.* Search for a two-Higgs-boson doublet using a simplified model in $p\bar{p}$ collisions at $\sqrt{s}$ = 1.96 TeV. *Phys.Rev.Lett.* **110**, 121801 (2013). [arXiv:1212.3837].

[11] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys.Lett.* **B716**, 1–29 (2012). [arXiv:1207.7214].

[12] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys.Lett.* **B716**, 30–61 (2012). [arXiv:1207.7235].

[13] Alwall, J. *et al.* MadGraph 5 : Going Beyond. *JHEP* **1106**, 128 (2011). [arXiv:1106.0522].

[14] Sjostrand, T. *et al.* PYTHIA 6.4 physics and manual. *JHEP* **05**, 026 (2006). [arXiv:hep-ph/0603175].

[15] Ovyn, S., Rouby, X. & Lemaitre, V. DELPHES, a framework for fast simulation of a generic collider experiment (2009). 0903.2225.

[16] The leptonic W boson is reconstructed by combining the lepton with neutrino, whose transverse momentum is deduced from the imbalance of momentum in the final state objects and whose rapidity is set to give $m_{\ell\nu}$ closest to $m_W = 80.4$ GeV.

[17] Bache, K. & Lichman, M. UCI machine learning repository (2013). URL `http://archive.ics.uci.edu/ml`.

[18] Cheng, H.-C. & Han, Z. Minimal Kinematic Constraints and m(T2). *JHEP* **0812**, 063 (2008). [arXiv:0810.5178].

[19] Barr, A., Lester, C. & Stephens, P. m(T2): The Truth behind the glamour. *J.Phys.* **G29**, 2343–2363 (2003). [arXiv:hep-ph/0304226].

[20] Rogan, C. Kinematical variables towards new dynamics at the LHC (2010). [arXiv:1006.2727].

[21] Buckley, M. R., Lykken, J. D., Rogan, C. & Spiropulu, M. Super-Razor and Searches for Sleptons and Charginos at the LHC (2013). [arXiv:1310.4827].

[22] Hocker, A. *et al.* TMVA - Toolkit for Multivariate Data Analysis. *PoS* **ACAT**, 040 (2007). [arXiv:physics/0703039].

[23] Bergstra, J. *et al.* Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)* (Austin, TX, 2010). Oral Presentation.

[24] Warde-Farley, D. *et al.* *pylearn2* (2011). Http://deeplearning.net/software/pylearn2.

Methods

In training the neural networks, the following hyperparameters were predetermined without optimization. Hidden units all used the *tanh* activation function. Weights were initialized from a normal distribution with zero mean and standard deviation 0.01, except in the first layer, which had standard deviation 0.1 (because there were relatively few inputs to this layer). Gradient computations were made on mini-batches of size 100. A momentum term increased linearly over the course of the first 200 epochs from 0.5 to 0.99 where it stabilized. The learning rate decayed by a factor of 1.0000002 every batch update until it reached a minimum of 10^{-6}.

The data sets were nearly balanced, with 53% positive examples in the HIGGS data set, and 46% positive examples in the SUSY data set. Input features were standardized over the entire train/test set with mean zero and standard deviation one, except for those features with values strictly greater than zero – these we scaled so that the mean value was one.

Computations were performed using machines with 16 Intel Xeon cores, an NVIDIA Tesla C2070 graphics processor, and 64 GB memory. All neural networks were trained using the GPU-accelerated Theano and Pylearn2 software libraries [23, 24].