nature COMMUNICATIONS

ARTICLE

Received 19 Feb 2014 | Accepted 4 Jun 2014 | Published 2 Jul 2014

DOI: 10.1038/ncomms5308

Searching for exotic particles in high-energy physics with deep learning

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

Collisions at high-energy particle colliders are a traditionally fruitful source of exotic particle discoveries. Finding these rare particles requires solving difficult signal-versus-background classification problems, hence machine-learning approaches are often used. Standard approaches have relied on 'shallow' machine-learning models that have a limited capacity to learn complex nonlinear functions of the inputs, and rely on a painstaking search through manually constructed nonlinear features. Progress on this problem has slowed, as a variety of techniques have shown equivalent performance. Recent advances in the field of deep learning make it possible to learn more complex functions and better discriminate between signal and background classes. Here, using benchmark data sets, we show that deep-learning methods need no manually constructed inputs and yet improve the classification metric by as much as 8% over the best current approaches. This demonstrates that deep-learning approaches can improve the power of collider searches for exotic particles.

[1] Department of Computer Science, UC Irvine, Irvine, California 92617, USA. [2] Department of Physics and Astronomy, UC Irvine, Irvine, California 92617, USA. Correspondence and requests for materials should be addressed to P.B. (email: pbaldi@uci.edu) or to D.W. (email: daniel@uci.edu).

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 antiprotons 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 have 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, the ratio of the sample likelihood functions in the two considered hypotheses, shown by Neyman and Pearson[2] to be the optimal discriminating quantity. 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.

The relative likelihood function is a complicated function in a high-dimensional space. Although any function can theoretically be represented by a 'shallow' classifier, such as a neural network with a single hidden layer[3], an intractable number of hidden units may be required. Circuit complexity theory tells us that deep neural networks (DN) 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[4,5]; the adjustments to the weights in the early layers of a DN 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 nonlinear functions of the input features that capture physical insights about the data. Although helpful, this approach is labour-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—for example, 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[6,7]; and (4) pre-training the initial layers of the network with unsupervised learning methods such as autoencoders[8,9]. This second approach attempts to learn a useful layered representation of the data without having to backpropagate through a DN; standard gradient descent is only used at the end to fine-tune the network. With these methods, it is becoming common to train DNs 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.

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.

Results

Particle collisions. The vast majority of particle collisions do not produce exotic particles. For example, though the Large Hadron Collider (LHC) 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 distinguishing collisions which produce particles of interest (signal) 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, θ (polar) and ϕ (azimuthal). For convenience, at hadron colliders, such as Tevatron and LHC, the pseudorapidity, defined as $\eta = -\ln(\tan(\theta/2))$ is used instead of the polar angle θ. 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 transverse to 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. The momentum imbalance in the longitudinal direction along the beam cannot be measured at hadron colliders, as the initial state momentum of the quarks is not known.

Benchmark case for Higgs bosons. The first benchmark classification task is to distinguish between a signal process where new theoretical Higgs bosons (HIGGS) 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[10] and the Tevatron colliders[11].

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[12] and CMS[13] experiments. The light Higgs boson decays predominantly to a pair of bottom quarks, giving the process:

$$gg \to H^0 \to W^\mp H^\pm \to W^\mp W^\pm h^0 \to W^\mp W^\pm b\bar{b}, \quad (1)$$

which leads to $W^\mp W^\pm b\bar{b}$, see Fig. 1. The background process, which mimics $W^\mp W^\pm b\bar{b}$ without the Higgs boson intermediate

Note that the leptonic W boson is reconstructed by combining the lepton with the 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.

Whereas 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.

Benchmark case for supersymmetry particles. 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[17–20] 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 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 (ref. 14) event generator assuming 8 TeV collisions of protons as at the latest run of the Large Hadron Collider, with showering and hadronization performed by PYTHIA[15] and detector response simulated by DELPHES[16]. 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.

The search for supersymmetric particles is a central piece of the scientific mission of the Large Hadron Collider. The strategy we

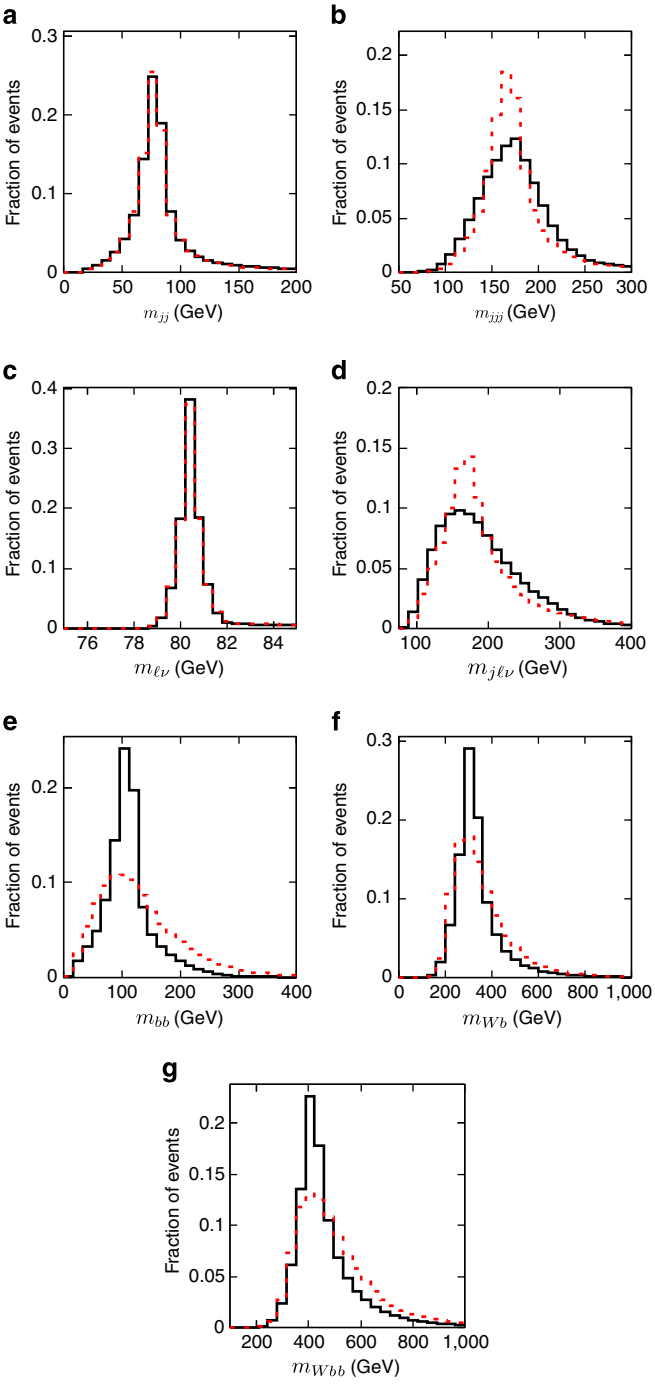

Figure 3 | High-level input features for Higgs benchmark. Distributions in simulation of invariant mass calculations in $\ell\nu jjb\bar{b}$ events for simulated signal (black) and background (red) events.

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 with one for the Higgs case). Instead, a great deal of intellectual energy has been spent in attempting to devise features that 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,

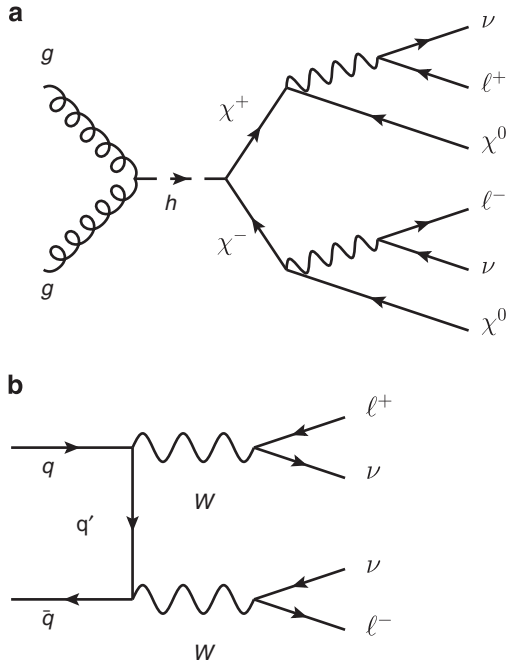

Figure 4 | Diagrams Example diagrams describing the signal process involving hypothetical supersymmetric particles $\chi^{\pm}$ and χ^0 along with charged leptons $\ell^{\pm}$ and neutrinos ν (**a**) and the background process involving W bosons (**b**). In both cases, the resulting observed particles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

- stransverse mass M_{T2}: estimating the mass of particles produced in pairs and decaying semi-invisibly[17,18],
- $\not{E}_T^{Rel}$: $\not{E}_T$ 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 (ref. 19),
- super-razor quantities β_{R+1}, $\cos(\theta_{R+1})$, $\Delta\phi_R^{\beta}$, M_{Δ}^R, M_R^T, and $\sqrt{\hat{s}_R}$ (ref. 20).

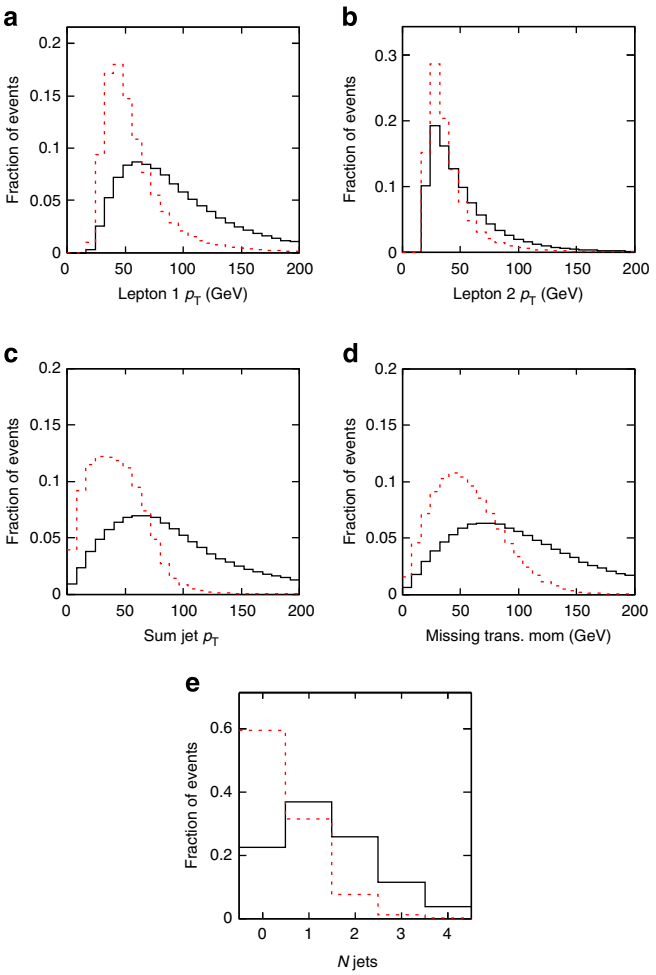

Figure 5 | Low-level input features Distribution of low-level features in simulated samples for the signal (black) and background (red) benchmark processes.

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 widely-used TMVA package[21], which provides a standardized implementation of common multivariate learning techniques and an excellent performance baseline.

Deep learning. We explored the use of DNs as a practical tool for applications in high-energy physics. Hyper-parameters were chosen using a subset of the HIGGS data consisting of 2.6 million training examples and 100,000 validation examples. Due to computational costs, this optimization was not thorough, but included combinations of the pre-training methods, network architectures, initial learning rates and regularization methods shown in Supplementary Table 3. We selected a five-layer neural network with 300 hidden units in each layer, a learning rate of 0.05, and a weight decay coefficient of 1×10^{-5}. Pre-training, extra hidden units and additional hidden layers significantly increased training time without noticeably increasing performance. To facilitate comparison, shallow neural networks were trained with the same hyper-parameters and the same number of units per hidden layer. Additional training details are provided in the Methods section below.

The hyper-parameter optimization was performed using the full set of HIGGS features. To investigate whether the neural networks were able to learn the discriminative information contained in the high-level features, we trained separate classifiers for each of the three feature sets described above: low-level, high-level and combined feature sets. For the benchmark, the networks were trained with the same hyper-parameters chosen for the HIGGS, as the data sets have similar characteristics and the hyper-parameter search is computationally expensive.

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 curves to illustrate the performance of the classifiers. Our primary metric for comparison is the area under the receiver operating characteristic curve (AUC), with larger AUC values indicating higher classification accuracy across a range of threshold choices.

This metric is insightful, as it is directly connected to classification accuracy, which is the quantity optimized for in training. In practice, physicists may be interested in other metrics, 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.

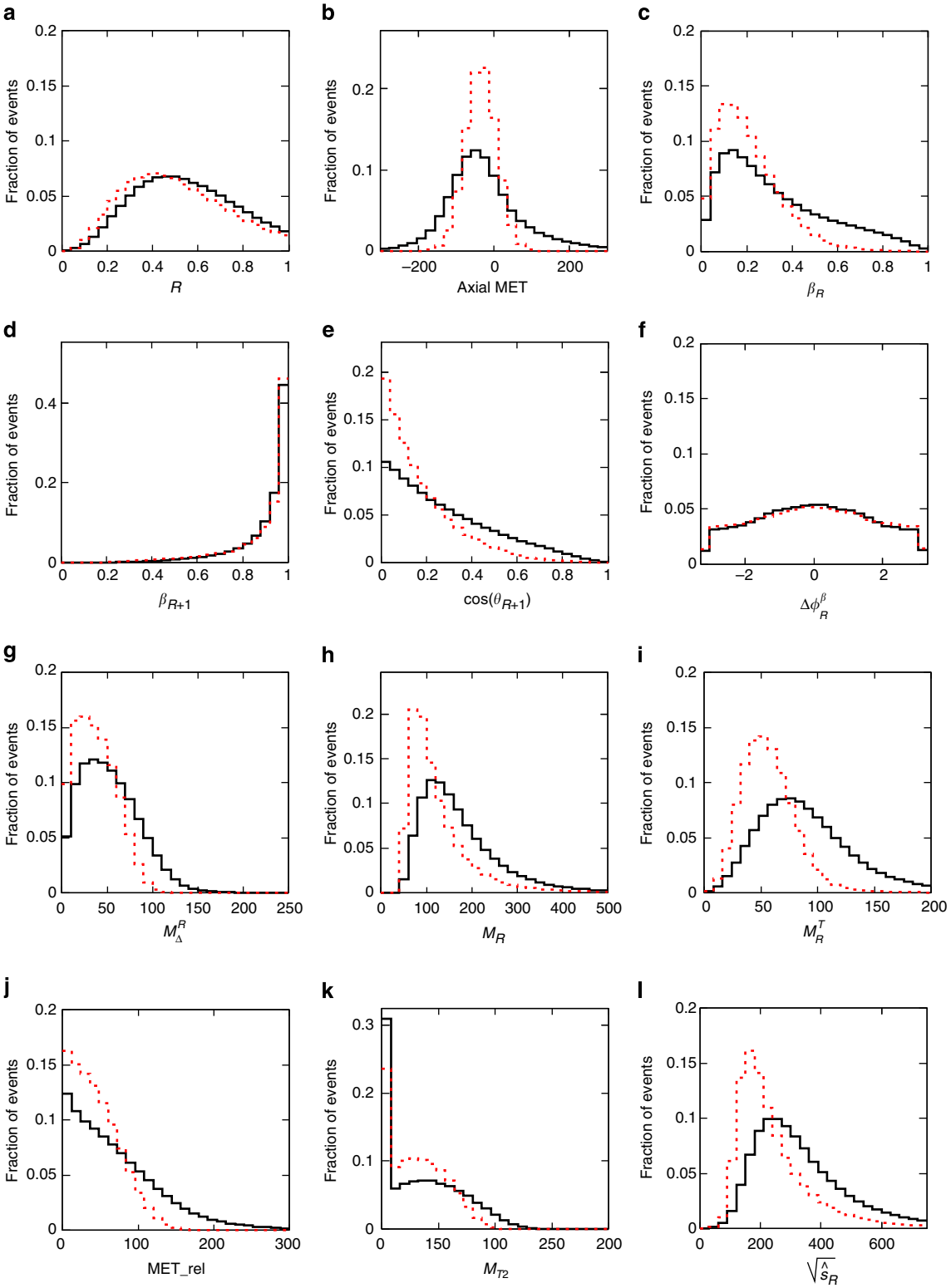

Figure 6 | High-level input features Distribution of high-level features in simulated samples for the signal (black) and background (red) benchmark processes.

In addition, we calculate discovery significance—the standard metric in high-energy physics—to demonstrate that small increases in AUC can represent significant enhancement in discovery significance.

Note, however, that in some applications the determining factor in the sensitivity to new exotic particles is determined not only by the discriminating power of the selection, but by the uncertainties in the background model itself. Some portions of the background model may be better understood than others, so that some simulated background collisions have larger associated systematic uncertainties than other collisions. This can transform the problem into one of reinforcement learning,

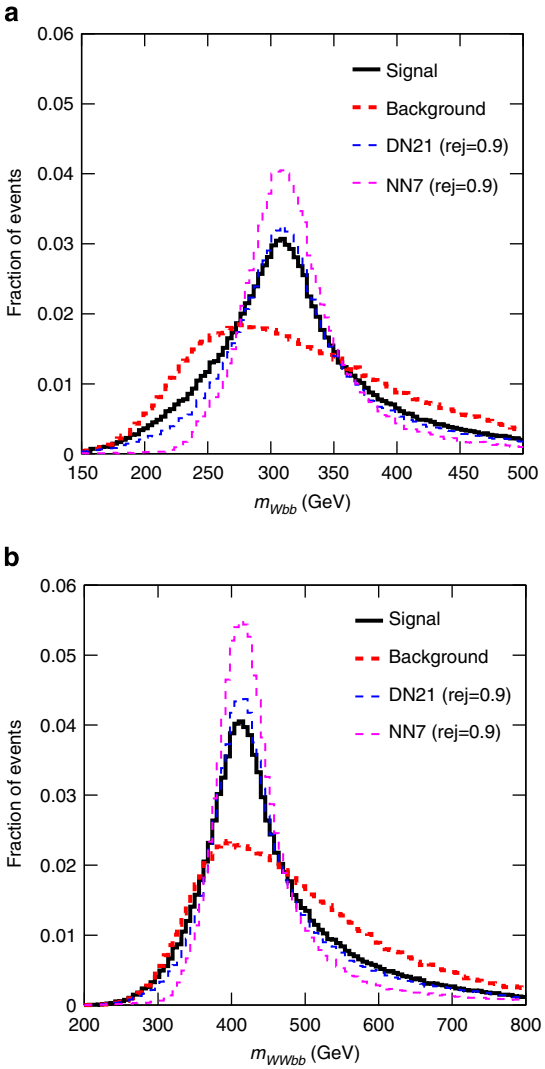

Figure 8 | Performance comparisons. Distribution of events for two rescaled input features: (**a**) m_{Wbb} and (**b**) m_{WWbb}. Shown are pure signal and background distributions, as well as events which pass a threshold requirement which gives a background rejection (rej) of 90% for a deep network with 21 low-level inputs (DN21) and a shallow network with seven high-level inputs (NN7).

units did very little to increase performance over a shallow network with only 30,001 parameters. Supplementary Table 5 compares the performance of the best shallow networks of each size with DNs of varying depth.

Although the primary advantage of DNs is their ability to automatically learn high-level features from the data, one can imagine facilitating this process by pre-training a neural network to compute a particular set of high-level features. As a proof of concept, we demonstrate how DNs can be trained to compute the high-level HIGGS features with a high degree of accuracy (Supplementary Table 6). Note that such a network could be used as a module within a larger neural network classifier.

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 nonlinear feature combinations to guide the shallow networks.

Our analysis shows that recent advances in deep-learning techniques may lift these limitations by automatically discovering powerful nonlinear feature combinations and providing better discrimination power than current classifiers—even when aided by manually constructed features. This appears to be the first such demonstration in a semi-realistic case.

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.

Table 2 | Performance comparison for the SUSY benchmark.

Technique	Low-level	High-level	Complete
AUC			
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)
$NN_{dropout}$	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)
$DN_{dropout}$	0.876 (<0.001)	0.869 (<0.001)	0.879 (<0.001)
Discovery significance			
NN	6.5σ	6.2σ	6.9σ
DN	7.5σ	7.3σ	7.6σ

BDT, boosted decision tree; DN, deep neural network; NN, shallow neural network; supersymmetry particle.
Each model was trained five times with different weight initializations. The mean area under the curve (AUC) is shown with s.d. in parentheses as well as the expected significance of a discovery (in units of Gaussian σ) for 100 signal events and 1,000 ± 50 background events.

Methods

Neural network training. In training the neural networks, the following hyper-parameters 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 0.1 in the first layer, 0.001 in the output layer and 0.05 in all other hidden layers. Gradient computations were made on mini-batches of size 100. A momentum term increased linearly over the first 200 epochs from 0.9–0.99, at which point it remained constant. The learning rate decayed by a factor of 1.0000002 every batch update until it reached a minimum of 10^{-6}. Training ended when the momentum had reached its maximum value and the minimum error on the validation set (500,000 examples) had not decreased by more than a factor of 0.00001 over 10 epochs. This early stopping prevented overfitting and resulted in each neural network being trained for 200–1,000 epochs.

Autoencoder pre-training was performed by training a stack of single-hidden-layer autoencoder networks as in ref. 9, then fine-tuning the full network using the class labels. Each autoencoder in the stack used tanh hidden units and linear outputs, and was trained with the same initialization scheme, learning algorithm and stopping parameters as in the fine-tuning stage. When training with dropout, we increased the learning rate decay factor to 1.0000003, and only ended training when the momentum had reached its maximum value and the error on the validation set had not decreased for 40 epochs.

Computation. 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[24,25]. Our code is available at https://github.com/uci-igb/higgs-susy.