CDF/ANAL/EXOTIC/GROUP/10282

Multivariate Photon ID for CDF

Jamie Ray[1], Craig Group[2,3], and Ray Culbertson[2]

[1] *Stanford University*

[2] *Fermi National Accelerator Laboratory*

[3] *University of Virginia*

(Dated: October 12, 2010)

Abstract

The potential to replace CDF's identification (ID) cuts for central photons with a multivariate ID technique is studied. Multivariate classifiers are trained with Monte Carlo (MC) samples to separate inclusive photons from photon-like jets (mostly π^0 's and η's). Although the current method of photon ID is effective, it excludes several powerful variables and ignores correlations. Multivariate techniques of photon ID are shown to provide significantly improved separation of true photons from the jet background. Accompanying Data/MC scale factors and plans to apply the new ID to the $h \to \gamma\gamma$ search are also reported.

Contents

I. INTRODUCTION

Fermilab's Tevatron collides protons and antiprotons at an energy of 1.96 TeV. These collisions can convert energy into mass in the form of the 17 fundamental particles – or combinations of them – predicted by the Standard Model (SM) of particle physics. Although it has been quite successful, there is now widespread agreement that the Standard Model is incomplete. As a result, the Tevatron is searching for new particles, such as the Higgs Boson, that can complete the SM. These particles often decay quickly and can't be observed directly – one must instead search for decay signatures of more common particles. To understand what happened in a high-energy collision, then, it is necessary to be able to reconstruct the particles in the final state.

The Collider Detector at Fermilab measures variables like position and momentum with great precision. A combination of several detector layers and online measurement algorithms generates a number of detector variables as output from each collision, or 'event.' However, a given particle's behavior can depend on its family, mass, spin, or flavour, among other variables. In an environment of such complexity, there is no single detector output that can uniquely identify a particle's type. Instead, a more sophisticated examination of the detector output is needed.

Photons, or γ particles, are ubiquitous and important to identify. However, they can be hard to distinguish from two 'backgrounds.' The first background, an electron, can mimic the characteristics of a photon and be mistaken as one - for this reason, we say that electrons can 'fake' photons. Similarly, a pair of photons can be created in quark jets, usually from the decay of a π^0 or η. In this case, two photons are *not* better than one, as the processes resulting in single photons are different from those creating pairs, and the distinction between such processes is necessary to truly understand an event. A successful method of photon identification (ID), then, should include as many true photons as possible while excluding a large fraction of these 'fakes.' The standard version of photon ID defines allowable ranges for the values of several detector variables, also known as 'cutting' on the values or using 'cuts.' An improvement in photon ID would result in better separation of the 'fake' background, synonymous with better sensitivity to photons. As such, it would lead to more powerful searches.

II. DETECTOR

The CDF detector has been extensively described in References [1, 2]. The variables relevant to photon identification include those concerning the detector's hadronic and electromagnetic calorimeters, as well as tracking information. Specifically, photons are uncharged, and so are expected to lack an associated track. The corresponding detector variables are the number and Pt of associated tracks. Furthermore, photons have a strong electromagnetic interaction, and are thus expected to deposit most of their energy in the electromagnetic (EM) calorimeter. In addition to the amount, this expected deposition has other defining characteristics such as shape. As a result, the variables pertinent to photon ID are those that either contain information about tracks measured near the photon, or energy deposition (calorimetry) in the photon's vicinity. CDF's standard photon ID uses cuts on a combination of several of these variables to restrict candidates to the expected signature, which lacks an associated track and has a well-understood EM calorimeter shower.

III. STANDARD PHOTON ID CUTS

Photon identification (ID) at CDF defines allowable ranges for several relevant variables within which a detected object can be classified as a photon. These ranges, or 'cuts,' are chosen to include a high proportion of true photons while rejecting most other particles. This standard cut-based ID has been used with minor modifications for several years, and also allows users to search for either central (found in the most sensitive region of the detector) or plug (located in the less-precise endcaps) photons. As a result of poorer understanding and detector resolution in the plug region, the majority of physics analyses are restricted to using only central photons. As a result, we focus on central photons, but note that the same methods can be applied to a multivariate 'plug' ID (see Appendix B). There is also a 'loose' version of the photon ID – in essence, a scaled version of the standard cuts – that can be used to include more photons. See Table I for a full description of the standard cuts.

Again, the two backgrounds that most commonly 'fake' a photon's detector signature are electrons, and jets from quark hadronization. The former are classified with photons as electromagnetic objects and interact almost identically in the electromagnetic calorimeter. However, electrons are charged and therefore create a track in the silicon tracker (SVX)

Variable	Standard Cut	Loose Cut
EtCorr	> 7 GeV	Same
CES X and Z Fiducial	Ces $\|X\| < 21$ cm, $9 <$ Ces $\|Z\| < 230$ cm	Same
HAD/EM	$< .125$ OR $< 0.055 + .00045$*EtCorr	$< .125$
Cone 0.4 IsoEtCorr	< 0.1*EtCorr for EtCorr ≤ 20	< 0.15*EtCorr for EtCorr ≤ 20
	ELSE: $< 2.0 + .02$*(EtCorr-20.0)	$<$ ELSE: $3.0 + .02$*(EtCorr-20.0)
χ^2 (Strips + Wires)/2.0	< 20	None
N track (N3D)	≤ 1	None
Track p_T	$< 1.0 + .005$*EtCorr GeV	$< .25$*EtCorr
Cone 0.4 Track Iso	$< 2.0 + .005$*EtCorr	< 5
2nd CES Cluster E*sin θ	< 0.14*EtCorr for EtCorr < 18	None
	$< 2.4 + 0.01$*EtCorr for EtCorr > 18	

TABLE I: The cut-based photon ID currently used for central photons at CDF. Includes a standard and loose version. For variable definitions see Ref. [3].

and/or central outer tracker (COT). Photons, being uncharged, will not have an associated track and can be distinguished from electrons in this way. In contrast, jets may contain short-lived particles that can decay to pairs of photons, or others that may themselves leave a similar signature. The former commonly include neutral pi mesons (π^0) and eta mesons (η), the decays of which are likely to include pairs of photons in the final state. These pairs are sometimes nearly collinear and barely separated, thus faking the detector response of a single photon. Other particles are able to fake the photon signature directly through the combination of a missing track with a large deposit in the electromagentic calorimeter. Thus, it is possible for jets to result in values of the cut variables that fall within the defined ranges. The frequency of jets faking photons has also been studied in Reference [4], suggesting contamination at the .1% level. With a high enough production of jets, this fake rate results in a background of photon-like jets that can threaten the purity of photon samples.

IV. BENEFITS OF MULTIVARIATE METHODS

In comparison with the standard process of identifying photons using independent cuts, new computing methods can use complex algorithms to train what is known as a 'classifier' to identify photons and reject background fakes. These techniques, in considering all variables combined instead of checking each value independently, are called 'multivariate' (MV). The process of creating a multivariate ID and its use differ from a more standard, cut-based ID. However, the cut-based system has two potential flaws that may cause it to underperform a multivariate method. Some trade-offs are discussed below:

- Cuts, if performed independently, cannot use the information of correlations between variables to discriminate between true and fake photons. For example, one might imagine that two variables correlate positively for photons but negatively for jets. Cuts simply yield a boolean result for each variable depending on whether it is in the allowed range – they include no other information. One can describe a cut-based ID in terms of the 'variable-space' – a region in the space of all variable combinations that is considered to contain true photons. The space for a positive photon ID using cuts is (hyper) rectangular. See Figure 1 for an illustration of the cut-space.

- Cuts suffer from arbitrariness – the exact endpoints of allowed ranges often lack real justification. Again, this type of system creates a 'black and white' photon ID with one region of variable-space that definitely contains photons, and another (the rest of the space) that does not. In reality, for some of these variables a midrange value is more representative of a 'true' photon than an extreme one. However, with the standard cuts a photon that has the most photon-like signature except for a single off-mark variable will still be excluded from the photon ID. In contrast, a multivariate method can weigh signal-like values of some variables to allow others to vary within a wider range.

- Both of the above methods can sculpt the traditional hyper-rectangular variable space defined by cuts into a smoother but less easily-visualized ID space. The multivariate method generates a single output value given all of the input detector variables. This value is continuous and can fall within some range, with one extreme defined as signal-like and the other background-like. Cuts also give an output value that is simply a

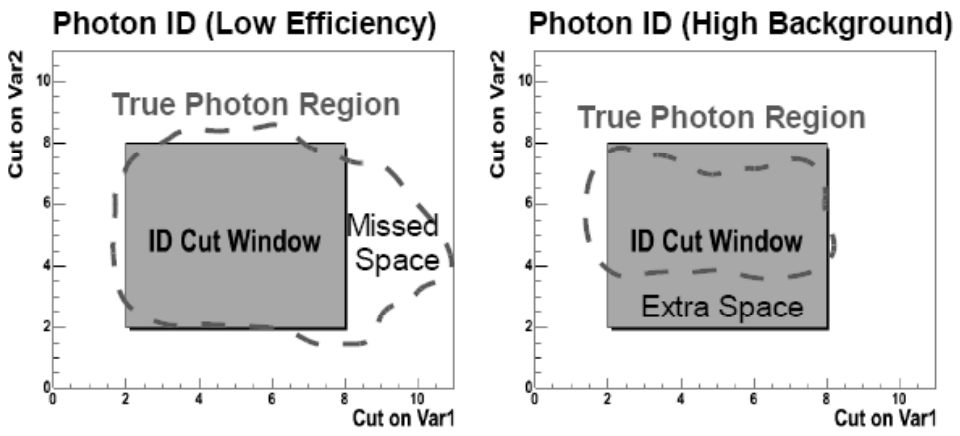

FIG. 1: Cartoon illustrating possible disadvantages to using a cut-based ID. Note that if there are only two variables (Var1 and Var2) the variable-space is 2D and the cut-space is rectangular. In reality, the variable-space is higher-dimensional and the cut-space is a hyper-rectangle. At left, the cut-space isn't shaped properly to include the extra photons with high Var1, resulting in a low-efficiency scenario. At right, it is too large and includes an extra space outside of the 'true photon region,' resulting in high background.

combination of the boolean value for each individual cut – the output is either true, or false. With a continuous output, users can choose how signal-like a particle must be for inclusion as a 'true' photon. This choice, then, defines one of many possible ID spaces that exist for all desired stringencies of photon ID. A cut-based ID, on the other hand, cannot be easily adjusted to be more or less stringent.

- Cuts are quick to assess and simple. Even if endpoints are arbitrary, the rough allowable ranges sketched by cuts can be justified in the context of expected photon behavior. A multivariate classifier may take longer to evaluate a photon, and the process of generating an output value is less transparent. The former concern is negligible for the methods discussed here, and the latter should not be cause for alarm if the ID is trained carefully and used prudently.

V. TRAINING SAMPLES

Multivariate techniques require a dedicated signal sample of particles known to be 'true' photons and a background sample representative of 'fake' photons. Through the use of iterative training algorithms, these techniques then 'learn' how to distinguish between the signal and background samples. To acheive the best performance, it is important to carefully choose training samples that match the desired use of the multivariate classifier. For the case of a general photon ID, a classifier should be able to positively identify a wide range of photons and be able to discriminate against all significant backgrounds. It is possible that there may be pathological cases of trying to ID a certain rare type of photon or remove a rare background signature, but in general it is desirable to use only one multivariate ID. As noted above, the most common photon 'fakes' are electrons, and neutral pions or etas from jets. However, electrons can in most cases be separated by examination of the tracking variables, with the only discernable difference between an electron and photon being the existence of an associated track. It is feasible that the detector might miss a track, or assign a track to a true photon, but in these instances there is little reason to believe that other variables could help to correct the ID process because both particles interact in the same way with the detector calorimetry. In contrast, jets, while lacking the separating potential of the electron track, can differ for various reasons from the detector signature of photons that they mimic. For very photon-like jets these differences are small, and thus they can't be easily separated by cuts, but a multivariate technique may recognize patterns in the ways that such jets diverge from the photon signature.

As the implementation of a new photon ID seems better suited to separating true photons from jets, the background training sample was chosen to consist of a sample of photon-like jets. It did not include electrons – instead, electrons can be discriminated against separately using the pertinent cuts from the current ID. All photons used for training were simulated by Monte Carlo (MC). The signal photons were chosen from a simulated sample of inclusive photon production [8], while 'fake' photons were chosen from a sample of simulated jets [9]. All selected 'photons' were required to pass a set of loose photon cuts (See Table I to ensure that they all looked reasonably like photons. This precaution forces the training of multivariate techniques to focus on the most difficult cases, assuming that the less 'photon-like' fakes will be easily separated anyway. Candidates also had to pass the relevant electron

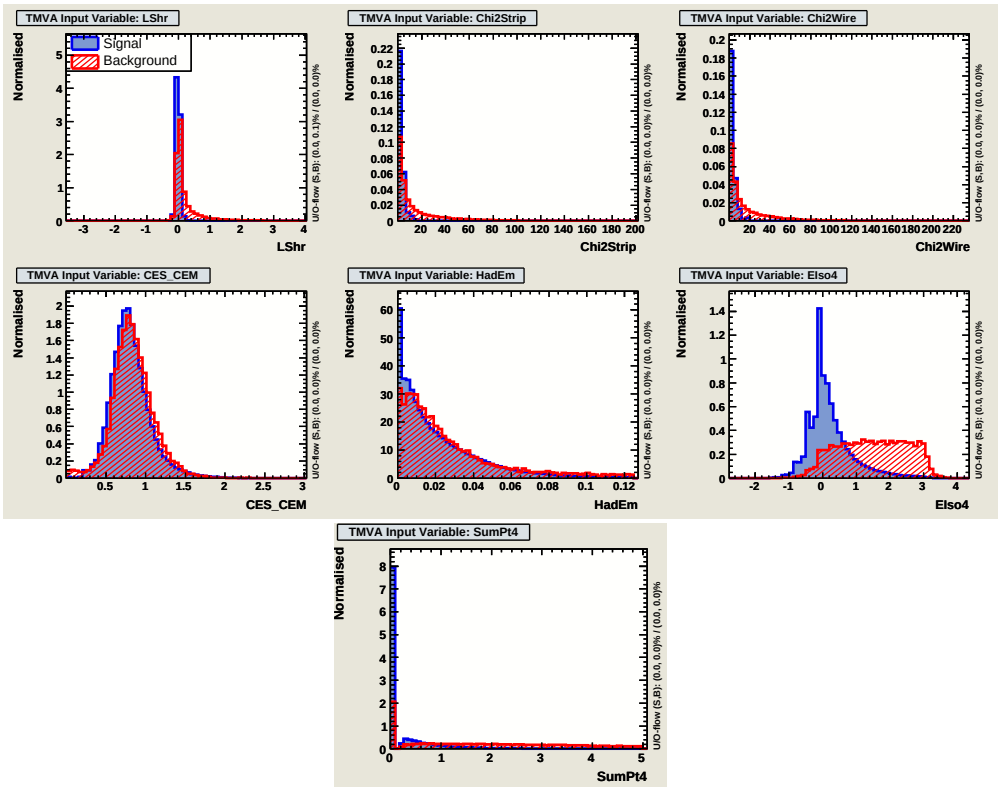

FIG. 2: Distributions of each of the input training variables for signal (blue) and background (red). Note that regions of high overlap indicate background that isn't easily separated with simple cuts.

rejection cuts, as it is assumed these cuts will be applied in conjunction with the multivariate ID.

A. Training Variables

As illustrated by CDF's standard photon ID, good central photons are expected to have:

- No associated track
- Good isolation - no tracks or energy in the same cone
- Most energy deposited in EM Cal
- EM energy in expected shape

Again, as electrons should be easy to separate through examination of tracking information, the classifiers are only trained against a background of jet 'fakes' and electron cuts

are kept separate. This means that variables useful for rejecting electrons – which are the number of tracks (N3D), the track P_T, and the 2nd CES cluster energies – aren't included in training (See training variables chosen below, and Figure 2 for the distributions of these variables in the signal and background samples). In particular, efforts were made to train with variables that have the same values for electrons and photons. Such care is valuable because it allows electrons to be substituted for photons when measuring the effectiveness of the multivariate ID. For example, electrons must be used to calculated scale factors that assess the ability of the detector simulation to model ID efficiencies. Given the similarity of calorimeter interactions between these two particles, all calorimetry variables should satisfy this requirement. However, while the Cone 0.4 Track Iso (SumPt4) variable contains tracking information, it is useful in discriminating against both electrons and jets. This is because it sums all of the tracks in a cone, and jets can easily include other particles besides the 'fake' photon that may create tracks. Thus, it *is* included in training. In order to use electrons in place of photons, one must simply remember to correct the SumPt4 by subtracting the P_T of the electron's track. See Section VIII and Reference [5] for details on electron substitution. The following detector variables relevant to central photons were chosen for training the multivariate ID methods:

- Had/Em (HadEm): This is a ratio of energies deposited in the hadronic (Had) and electromagnetic (EM) calorimeters. Chosen because it is used in the standard cuts. Photons are expected to deposit mostly in the EM calorimeter.

- Cone 0.4 IsoEtCorr (EIso4): The sum of additional energy in a cone with radius .4 around the photon. Taken from standard cuts. Photons are expected to be 'isolated' without extra particles in the same cone.

- Chi2 (Strips+Wires): Compares the electromagnetic shower to the expected lateral shape. Taken from standard cuts, *except* that we use Strip and Wire values separately so that e.g. a bad Wire shape doesn't throw out a photon.

- Cone .4 Track Iso (SumPt4): The sum of track momenta in a cone (radius .4) around the photon. Taken from standard cuts. Photons are again expected to be 'isolated' without extra particles in same cone. Technically a 'tracking variable' but useful for electrons *and* jets as both can increase SumPt4.

- LShr: Also uses lateral shower shape, but compares the lateral sharing of energy *between* towers to expectations. Not a standard cut variable, but is used often for electrons and gives additional information about whether energy deposition is 'photon-like' as photons are expected to have little sharing.

- CES/CEM Energy (CES_CEM): The ratio of the energy from the shower maximum detector to the total measured energy. Not a standard photon cut ID variable, but should provide some separation from π^0 's.

B. Simulation Particle Matching

The input training samples were processed further to ensure that they accurately represented the cases that a photon ID classifier would be required to separate. Significantly, the background sample can contain events in which true photons, not just jets that fake a photon signature, are produced. These most often take the form of initial-state radiation, or ISR, from a quark in the underlying event. Of course, being real photons, it makes sense that these particles would pass the loose requirements described above. However, it is important that they are not included in the background sample; otherwise, the multivariate classifier will be trained to accept some photons and reject others. It is clearly undesirable to train the classifier to reject any true photon, so the background candidates are checked to ensure that they aren't actually photons radiated by a quark. This requires a matching between the generator-level particles in the simulation and the particles reconstructed from the simulated detector output. As pions and etas are known to be the most common fakes found in jets, a simple approach might require that all candidate background photons be matched in this way to a pion or eta at the generator level. However, a more general background sample would include other possibilities for jets faking photons. As a result, the matching process required that any candidate background photon could not match to a generator level radiated photon. The 'extra' jets gained through this choice of matching were found to comprise around 25% of the total selected sample. A validation of the extra candidates reveals that their distributions for the training variables are often less signal-like than the most common π^0 's or η's (See Figure 3), and therefore they should certainly be trained against.

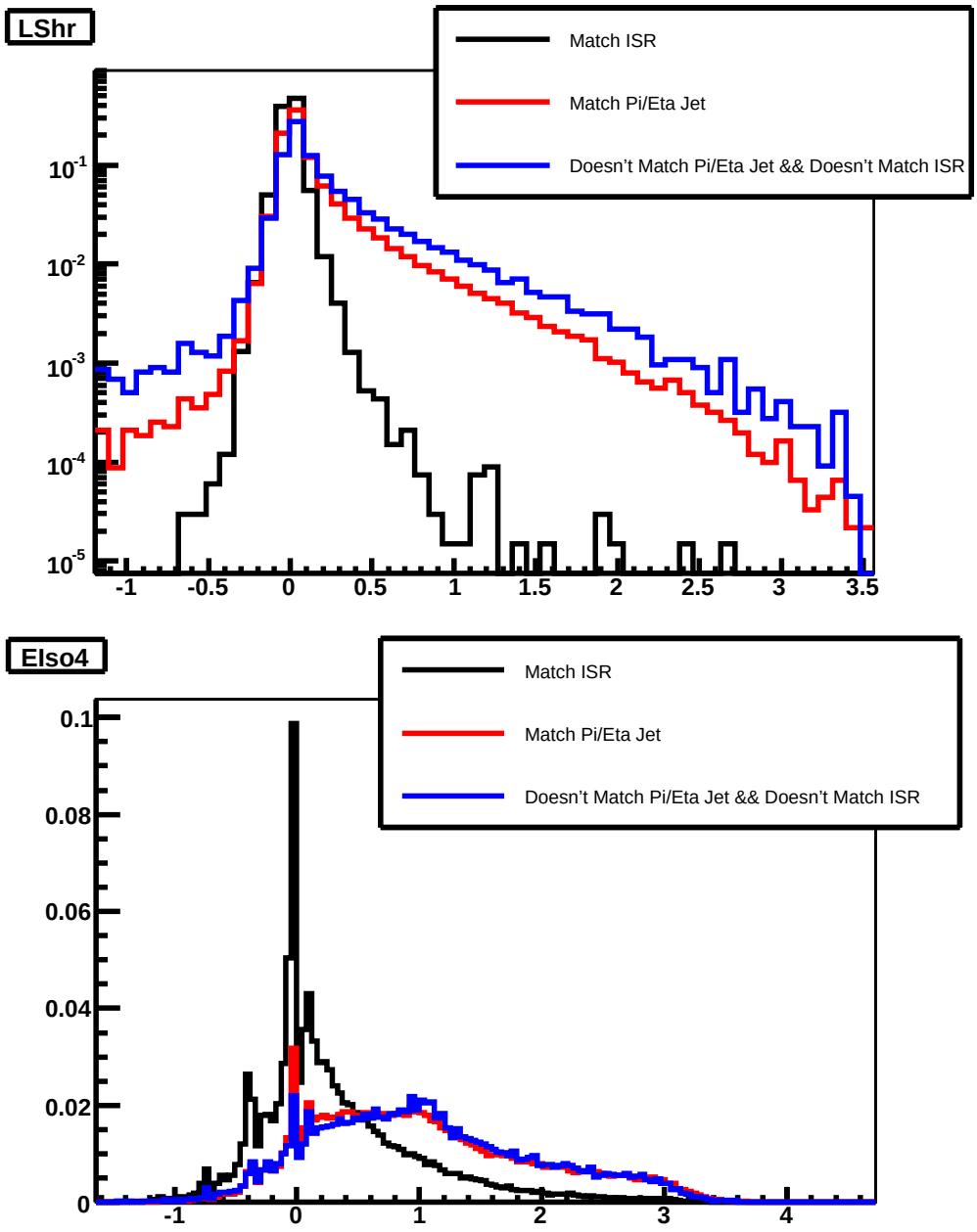

FIG. 3: Distributions of two representative training variables (LShr and EIso4) drawn for each of the different simulation particle-matching options. Similar distributions were drawn for all of the training variables, indicating in all cases that making the looser anti-ISR match includes background-like candidates.

C. Energy Matching

Finally, the energy distribution – specifically, the energy directed in a plane transverse to the beam direction (E_T) – of the background fakes was observed to differ from that of the signal photons. This is expected, as each type of event will have its own characteristic transverse energies. True photons, as shown in Figure 4, generally have a higher E_T than jet fakes. In this way, E_T may seem to be an ideal candidate for a cut that would keep signal photons but reject most of the fakes. While the standard version of the cuts does indeed require a minimum Et, as do the loose requirements made on the training photons, it is undesirable to train with such different energy distributions for two reasons. The first is that transverse energy is expected to correlate with the other variables relevant to photon ID, so these distributions in turn will be different and the multivariate technique will not be challenged. This relates to the second problem, which is that for generality the ID should be useful for both low-E_T and high-E_T photons, both of which are still plagued by a background with the same E_T. While the background distribution falls off more steeply at high E_T, it can still overwhelm the number of 'true' photons if produced more frequently. To force the classifier to discriminate between similar-energy signal and fake photons – and thereby prevent it from generating a simple cut on E_T – each of the fakes in the background sample was assigned a weight that would be visible during training. For example, a particle with a weight of three would be considered thrice as important to exclude from ID as a particle with unit weight. To prohibit unexpected behavior that may result from only having a few candidates to train with, the samples were limited to those ranges of E_T containing enough signal and fake photons to provide a representative sample with good statistics. In this case, we exclude ranges of E_T in which the weights are extremely high or the weights are 0, because either case indicates that the number of background 'fakes' is limited (See Figure 4). Even with this limitation, however, it is believed that the general signatures of the signal and background should be somewhat smooth so that the patterns recognized by a classifier can also be applied to photon ID at other energies.

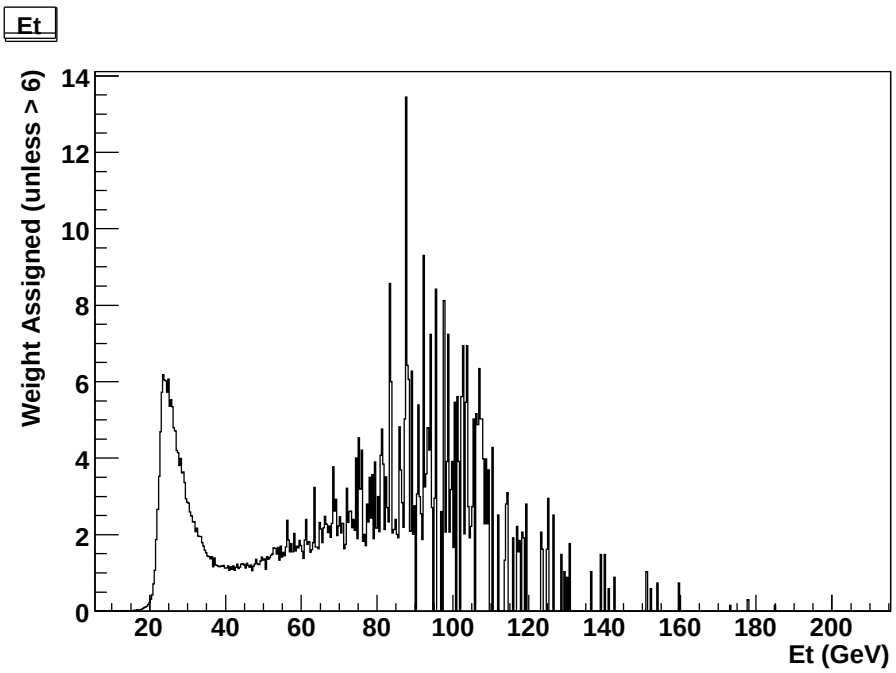

FIG. 4: Weights applied to background 'fake' candidates based on their E_T. After this weighting, the distributions of signal and background are seen to be matched in the ranges where weighting is valid. In the high-energy tail, the background distribution falls off more rapidly than signal, resulting in high computed weights. To avoid assigning large weights to this small number of events, training candidates were required to have E_T in ranges containing enough background candidates.

VI. TRAINING PROCESS

The seven relevant variables were stored for each of the MC simulated 'photons,' which were required to pass the above selection criteria. To summarize, this preprocessing of the simulated samples consisted of the following steps:

- Apply 'loose' cuts from Table I
- Apply 'standard' electron tracking cuts (on N3D, Track P_T, and 2nd CES Cluster) from Table I
- Make generator-level matching requirements so that 'fake' candidates aren't real photons radiated from quarks
- Assign weights to 'fake' background based on E_T

Afterwards, the signal and background samples were ready for use. The resulting sample consisted of 764585 signal photons and 113309 background 'photons'. Multivariate training was accomplished using the Tools for Multivariate Analysis (TMVA) software package [6], available alone and with recent releases of the ROOT libraries. TMVA contains a variety of training methods, documented in the User's Guide. Initial training runs compared several multivariate algorithms for their performance using a small fraction of the signal and background samples. Figure 5 shows a comparison of these various classifiers using the standard benchmark of a signal efficiency vs. background rejection curve. The MLP (which is an artificial neural network, or NN) and Likelihood methods were found to perform the best in separating the reduced samples, with the added advantage of both training and evaluating photons quickly. It is assumed that good performance on a reduced training sample should be maintained when training with the full signal and background datasets. These two classifiers were then trained with the full samples. A full training is time-consuming, and splits the samples randomly with half of each sample being used to train, and the other half being used later to assess the performance of the trained classifier. This process helps to address the concern of 'overtraining' in which a multivariate method iterates so much that it is only useful for separating the exact samples it is trained with. Performance was found to be similar on both the training and test samples, suggesting that the classifiers were not overtrained (see Figure 6).

VII. RESULTING CLASSIFIERS

The performance of the resulting, fully-trained classifiers is shown in Figure 7. The MLP and Likelihood methods seem to have roughly equivalent performance, especially in the region of high ($> 80\%$) signal efficiency that is most likely to be useful. The performance of the cut-based ID on the same sample is shown as a black star, and is seen to significantly underperform the multivariate techniques. The comparison between these three ID possibilities is summarized in Table II. The signal efficiencies and background rejections were evaluated for the full samples described above, with the same weighting and requirement of electron tracking cuts, but without the range requirements for E_T. As such, performance is assessed for photons of any E_T, and the classifiers are still shown to separate this full sample well even though they were only trained for specific ranges, supporting the assumption

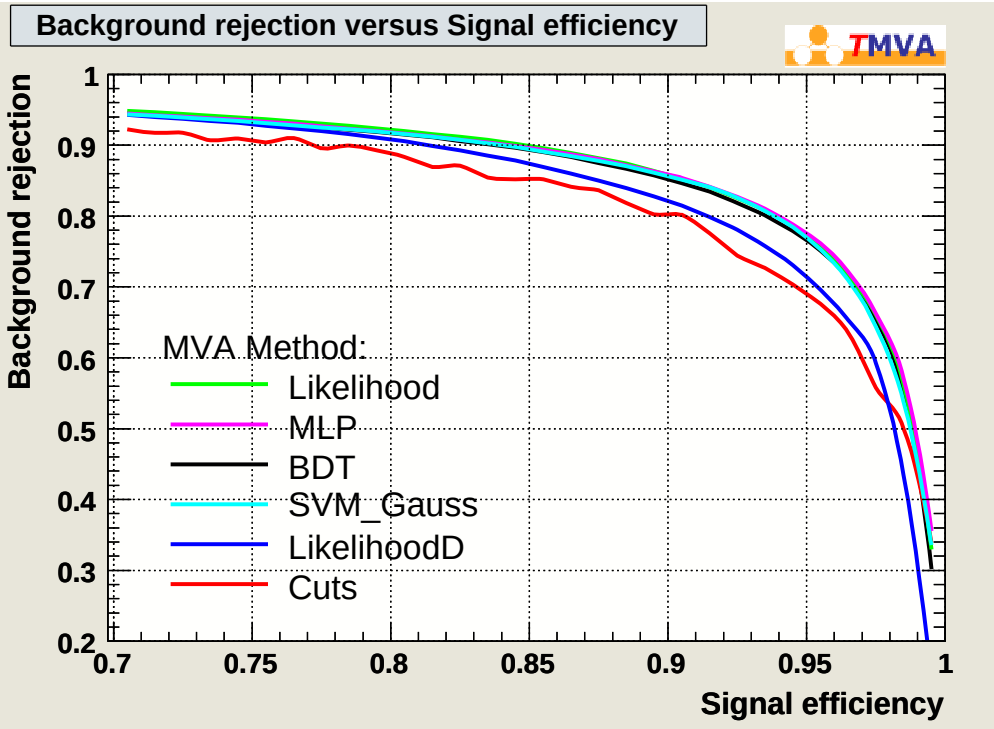

FIG. 5: Performance evaluation plot created by TMVA to assess signal efficiency and background rejection of all trained classifiers. In this step, the performance of several of the highest-recommended multivariate tools was tested on a subset of the signal and background samples. Note that MLP (the recommended Neural Net) and Likelihood acheive the highest efficiencies for a given background rejection, indicating superior performance.

that their use can be extrapolated to other values of E_T. It is important to note that the gains in efficiency or rejection made possible by the use of this multivariate ID are probably overestimated. The actual benefit depends on the proportion of jets expected to comprise the background for a given analysis, but we note that this will be a substantial fraction in most cases. As the multivariate ID is meant to separate photons from a jet background only, it should be combined with the standard tracking cuts (made in the sample preprocessing) used to reject the electron background. As the variables used for training are expected to be the same for photons and electrons, the combination of a trained multivariate technique with the electron cuts should be able to reject the electron background as well as the standard cut-based ID (for the same signal efficiency).

ID Method	Signal Efficiency	Background Rejection	Cut Value
Standard Cuts	91.6%	65.6%	All Cuts Applied
Neural Net (MLP)	91.6%	86.6%	0.868
Likelihood	91.6%	84.9%	-0.022
Neural Net (MLP)	99.0%	65.6%	0.396
Likelihood	98.4%	65.6%	-0.248

TABLE II: Comparing the performance of standard cuts and newly-trained multivariate ID methods. We assess the potential gains in signal efficiency with background rejection maintained at current (cut-based) levels, and gains in background rejection with signal efficiency similarly maintained. Improved signal efficiency of 10% or background rejection of 32% is acheived for the samples used above. Actual performance depends on signal and background composition and may be optimized at some other point. If optimization isn't feasible, choosing one of the above cuts to improve only signal efficiency or background rejection is recommended.

VIII. SCALE FACTORS

Finally, it is important to assess the extent to which simulated photons might differ from actual data to be able to extrapolate from the signal efficiencies for MC to performance on real photons in the data. For example, it is possible that the simulated detector response for a particular variable doesn't match the actual behavior of the detector. This kind of difference could lead to either artificially increased or decreased separability of true and fake photons, depending on how the inaccuracy of the simulation changes the variable distributions. The comparison between MC simulation and data is evaluated with what is known as a scale factor (SF), or ratio of the efficiencies for equivalent samples in data and simulation. A pure sample of photons in data is needed to be able to evaluate the true efficiency. However, such a sample isn't readily available, so instead a pure sample of electrons from the Z boson decay is used. This substitution is justified by the equivalent calorimetry for electrons and photons. It has been used extensively in the past [5] to study scale factors for the standard photon ID, and revealed that the simulated detector response results in an overestimation of the actual ID efficiency, with the data efficiency at around 95% of the efficiency for simulated electrons. In this case, the photon ID is considered to be the combination of the standard tracking

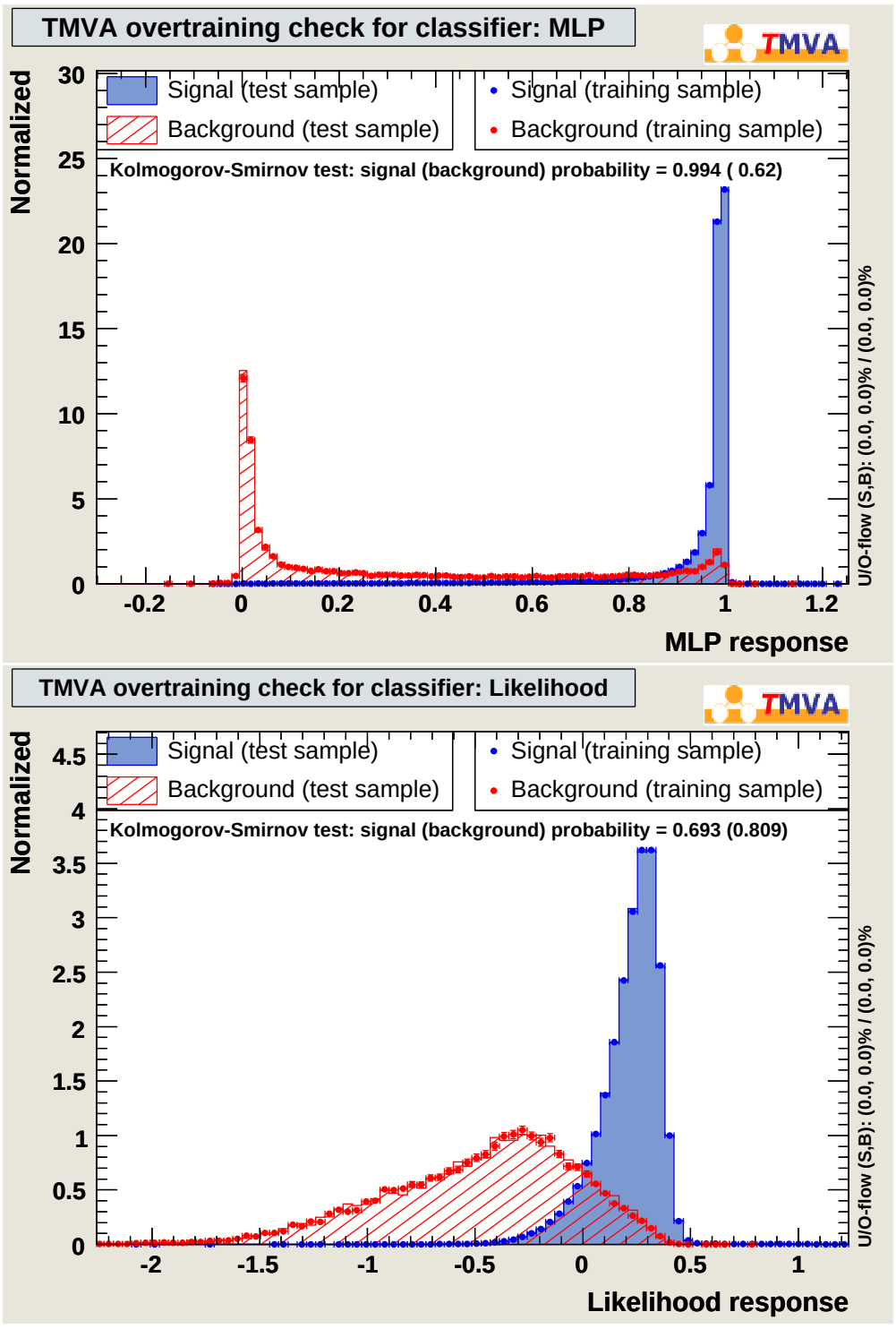

FIG. 6: Check against overtraining created by TMVA. Classifier outputs are plotted for training photons, and these distributions are compared with those for testing photons. Close agreement in output distributions suggests that multivariate methods don't use statistical fluctuations in the training sample and can discriminate other samples as effectively.

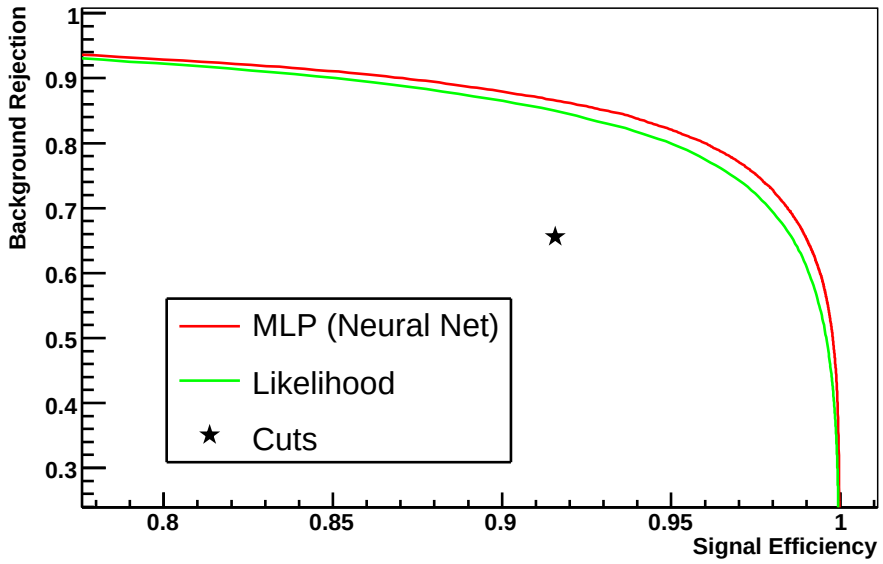

FIG. 7: Signal Efficiency vs. Background Rejection plot to measure performance of final, fully-trained classifiers. The efficiency and rejection of the standard photon cuts is shown as a black star for comparison.

cuts used to reject electrons, with the new multivariate technique used to reject jets. As in Reference [5], the denominators for these efficiencies are taken to be the electrons passing the standard loose cuts, and it must be stressed that these loose cuts should always be applied before using the multivariate output, as the classifiers were trained with 'loose' photons. A more complete explanation of the scale factor process can be found in Reference [5], but the same methods are applied here to calculate scale factors for the new multivariate classifiers. However, whereas the rigidity of the standard cut-based ID results in a unique efficiency for a given sample of photons, the multivariate ID allows the user to choose a desired efficiency by varying the output value required for positive identification. As a result, we calculate efficiencies for each number of vertices in the event and for a finely-binned array of possible cuts on the classifier output. An assortment of these efficiencies for different MLP working points are given in Table III. Users can then choose a desired stringency of photon ID and, using the methods described for weighting by number of vertices in Ref. [5], easily calculate a scale factor applicable to the analysis at hand.

MPL cut	sample	N vertex				
		1	2	3	4	5
	MC	0.961	0.946	0.927	0.908	0.893
0.05	≤p18	0.95	0.932	0.907	0.887	0.83
	>p18	0.95	0.913	0.893	0.87	0.82
	MC	0.956	0.942	0.921	0.902	0.887
0.15	≤p18	0.943	0.924	0.901	0.877	0.821
	>p18	0.943	0.905	0.885	0.861	0.805
	MC	0.95	0.936	0.915	0.896	0.881
0.25	≤p18	0.935	0.917	0.893	0.866	0.805
	>p18	0.934	0.897	0.879	0.848	0.795
	MC	0.945	0.931	0.909	0.888	0.875
0.35	≤p18	0.926	0.909	0.885	0.857	0.802
	>p18	0.926	0.889	0.871	0.837	0.795
	MC	0.937	0.923	0.9	0.88	0.866
0.45	≤p18	0.917	0.899	0.876	0.851	0.791
	>p18	0.92	0.878	0.86	0.82	0.778
	MC	0.927	0.912	0.889	0.869	0.855
0.55	≤p18	0.904	0.888	0.864	0.832	0.777
	>p18	0.905	0.861	0.843	0.802	0.771
	MC	0.915	0.9	0.877	0.856	0.839
0.65	≤p18	0.891	0.875	0.85	0.817	0.761
	>p18	0.893	0.846	0.827	0.788	0.747
	MC	0.893	0.878	0.855	0.834	0.811
0.75	≤p18	0.868	0.853	0.825	0.799	0.729
	>p18	0.865	0.821	0.798	0.749	0.73
	MC	0.857	0.84	0.817	0.798	0.77
0.85	≤p18	0.835	0.813	0.79	0.766	0.699
	>p18	0.828	0.779	0.751	0.704	0.697
	MC	0.753	0.742	0.723	0.711	0.691
0.95	≤p18	0.739	0.717	0.7	0.662	0.598
	>p18	0.724	0.679	0.65	0.583	0.596

TABLE III: MC and data efficiencies as calculated for the MLP using the tag and probe method. Efficiencies may also be available for additional MLP working points, and the Likelihood by contacting the authors.

IX. APPLYING THE MULTIVARIATE ID

After training, TMVA generates a standalone class for each of the trained classifiers. While it is recommended that the classifier output be evaluated within the framework of the TMVA Reader class, these standalone classes make it possible to export the multivariate ID

for use without the TMVA package. The output classes are contained in a single .C file and, as they are generated automatically, can be difficult to read or include very long statements. As a result, the output classes were cleaned up and modified into singleton classes, as their function can never change between instances anyway. An interface between these singleton classes and the standard CDF photon definitions was developed and will be implemented in the CDF code base (Stntuple), so that users can easily obtain the output value of the desired classifier for any photon object. Again, to be able to discriminate against an electron background one must make the cuts on tracking variables separately. Furthermore, the classifiers expect variables to fall within the ranges specified by the training samples, which had passed the standard loose cuts. As a result, the full recommended photon ID replaces the standard cuts with a combination of three steps, also shown in Table IV:

Variable	Cut
EtCorr	> 7
CES X and Z Fiducial	Ces $\|X\| < 21$ cm, $9 <$ Ces $\|Z\| < 230$ cm
HAD/EM	$< .125$
Cone 0.4 IsoEtCorr	< 0.15*EtCorr for EtCorr ≤ 20
	$< 3.0 + .02$*(EtCorr-20.0)for EtCorr > 20
Cone 0.4 Track Iso	$< .25$*EtCorr
N track (N3D)	≤ 1
Track p_T	N3D > 0: $< 1.0 + .005$*EtCorr GeV
2nd CES Cluster E*sin θ	< 0.14*EtCorr for EtCorr < 18
	$< 2.4 + 0.01$*EtCorr for EtCorr > 18
Multivariate Output	$>$ Value giving desired stringency

TABLE IV: Set of recommended cuts in order: loose photon cuts, standard electron-specific cuts, cut on multivariate output

- Apply standard set of loose photon cuts
- Apply electron-specific cuts
- Apply cut on multivariate output for desired efficiency and rejection

Again, this combination is as simple to use as the current photon ID cuts, while promising greater performance in scenarios with significant jet backgrounds, and similar performance for electrons. In order to optimally choose the cutoff (stringency) for the multivarate ID, a user will need MC samples containing signal and background events that can be evaluated for the classifier output. This will facilitate the creation of a signal efficiency and background rejection curve that is specific to the analysis, and the optimal point on this curve can be chosen with reference to the numbers of signal and background events expected before ID requirements (See the Appendix for an example optimization). If such optimization isn't possible, we recommend choosing one of two standard points from Table II – increasing efficiency while maintaining the background rejection of the current cuts, or increasing rejection to keep the efficiency constant.

X. CONCLUSIONS

In conclusion, the potential for replacing part of CDF's current photon ID with a more effective system has been investigated. While the electron background is expected to differ significantly in terms of tracking variables – suggesting that cuts can be effective – the background of jets faking a photon signature may not be effectively rejected with the standard cut-based ID. Instead, a multivariate technique seems to be a more powerful tool to separate photons from jet 'fakes' of all kinds, not only the most common π^0 's and η's. The use of a multivariate ID, in combination with the standard tracking cuts used to reject electrons, should outperform the cut-based ID in a variety of scenarios. This multivariate classifier is simple to use and more flexible, allowing users to choose the optimal stringency for the analysis at hand rather than accept a set efficiency and background rejection from cuts. The use of simulation in training this new ID is shown to be reasonable by evaluating scale factors between MC simulation and data, which seem to be quite close to the scale factor on the performance of the current cut-based ID. Combining ease of use and significant potential gains, the use of a multivariate method for identifying photons is recommended for any analysis at CDF that aims to separate true photons from a background containing jet fakes.

APPENDIX A: APPLICATION OF NEW MULTIVARIATE PHOTON ID TO THE SEARCH FOR $h \to \gamma\gamma$

CDF recently completed a search for the Higgs Boson in the diphoton decay channel ($h \to \gamma\gamma$, [7]). The resulting limits were added to the Fermilab's combined Higgs limits, contributing at the level of roughly 20 times the Standard Model in the low-mass regime. As this search requires the identification of *two* photons in each event and has a low number of expected signal events given a low branching fraction to photons, improvements in photon ID efficiency can result in significant gains. For this reason, the multivariate photon ID described here will be implemented in the next iteration of the Higgs search, completely replacing the standard central ID cuts. The cut on multivariate output was optimized for a Higgs mass of 120 GeV/c^2 using a rough calculation of significance for various values of ID efficiency and background rejection. The background for this decay is assumed to be comprised of 75% 'fake' photons in the form of jets, with the other 25% being real photons and thus an 'irreducible' background in the sense that they will have the same ID efficiency as photons from the Higgs decay. Using this background model, the significance (number of signal events divided by the square root of the number of background events) was computed for a number of possible cuts on the classifier output. The numerator, or signal, was taken to be the number of expected events multiplied by the photon ID efficiency squared (as two photons must pass the ID for each event). The background consists of three different types of events: the first has two true photons produced by background processes, the second includes one real and one fake (jet-based) photon, and the third has two fake photons. Fakes are a 'reducible' background, whereas true photons from other processes are 'irreducible' because they have the same ID efficiency as the photons from the Higgs decay. The total number of events observed in the Higgs mass window is around 300, for the test mass of 120 GeV used in the optimization. Preliminary results assumed a background comprised of 75% fake-fake or doubly-reducible events and 25% true-true or irreducible events. This is a naive estimate that ignores the contribution from true-fake, singly-reducible background events. The expected significance was then calculated using the multivariate ID efficiencies for true and 'fake' photons. Optimal significance results at a cut of 0.74 (See the maximum significance point in Figure 8) as the minimum allowed value of the neural net output, with a signal efficiency of 95.5% and background rejection of 81.3% (compared with 91.6 and

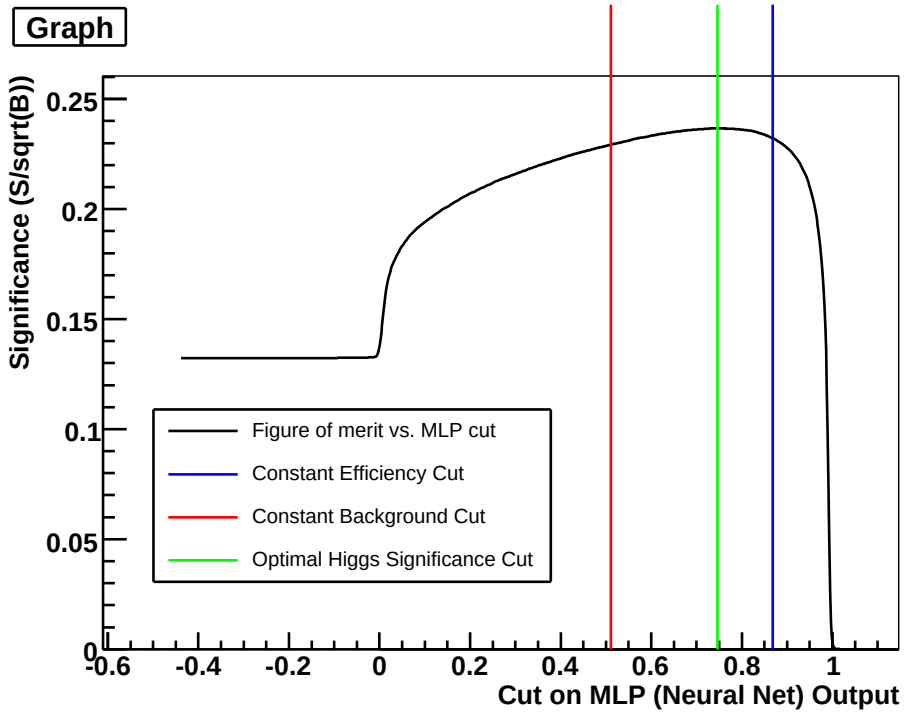

FIG. 8: Significance for the Higgs signal at 120 GeV/c^2 as a function of photon ID stringency (cut on Neural Net output). The optimal point appears at a cut of around .74 but a relatively flat plateau is observed, suggesting that a cut maintaining 'standard' efficiency *or* background rejection would also be a fine choice. A similar optimization can be conducted for any analysis with simulated signal and background samples, or one may use the training samples in order to calculate signal efficiencies and background rejections.

70.6 for standard cuts). This rough optimization predicts that an increase in significance of about 10% is possible by implementing the new multivariate ID. Indeed, it has already been tested at a Higgs mass of 120 GeV, and improved limits from 19.5 to 17.9 times the Standard Model prediction, a gain of 9%. Further optimization using a more sophisticated estimate of the background is possible and will be investigated when the multivariate ID is fully integrated into the Higgs search.

APPENDIX B: MULTIVARIATE ID FOR PLUG PHOTONS

The use of a multivariate ID technique for plug photons has also been investigated using the same approach applied for central photons. As such, it is possible that it could be optimized further based on differences between the central and plug detectors. The same simulated MC datasets were used, with similar matching and weighting preprocessing algorithms. Candidates were again required to pass the loose version of the cut-based ID for plug photons, in Table V. The variables chosen include those used in the standard cuts as well as some additional relevant variables that describe energ deposition. These added variables include a ratio of PES energy to total energy (PES_PEM), the PES Delta R, and splitting PES 5x9 into its component parts, 5x9U and 5x9V (See similar split for CES Strips and Wires). It is important to note that there are no separate 'electron variables' for plug photon ID, as tracking information is not used in the standard cuts with the exception of SumPt4, which has been discussed above. The trained multivariate classifiers again suggest that an improvement in ID performance is possible, subject to the same considerations as the central case. While plug photons are less frequently used, the Higgs diphoton analysis may include plug photons, for which an improved plug photon ID should be valuable. See Table VI and Figure 9 for more information on classifier performance.

Variable	Standard Cut	Loose Cut
PES U and V Fiducial	$1.2 < \|\eta\| < 2.8$	Same
HAD/EM	< 0.05 for ECorr ≤ 100 GeV	$< .125$
	ELSE: $< 0.05 + 0.026$*ln(ECorr/100)	
Cone 0.4 IsoEtCorr	< 0.1*EtCorr for EtCorr ≤ 20	< 0.15*EtCorr for EtCorr ≤ 20
	ELSE: $< 2.0 + .02$*(EtCorr-20.0)	ELSE: $< 3.0 + .02$*(EtCorr-20.0)
PEM χ^2	< 10	None
PES 5x9	> 0.65	None
Cone 0.4 track Iso	$< 2.0 + 0.005$*EtCorr	< 5

TABLE V: The cut-based photon ID currently used for plug photons at CDF. Includes a standard and loose version. For variable definitions see Ref. [3].

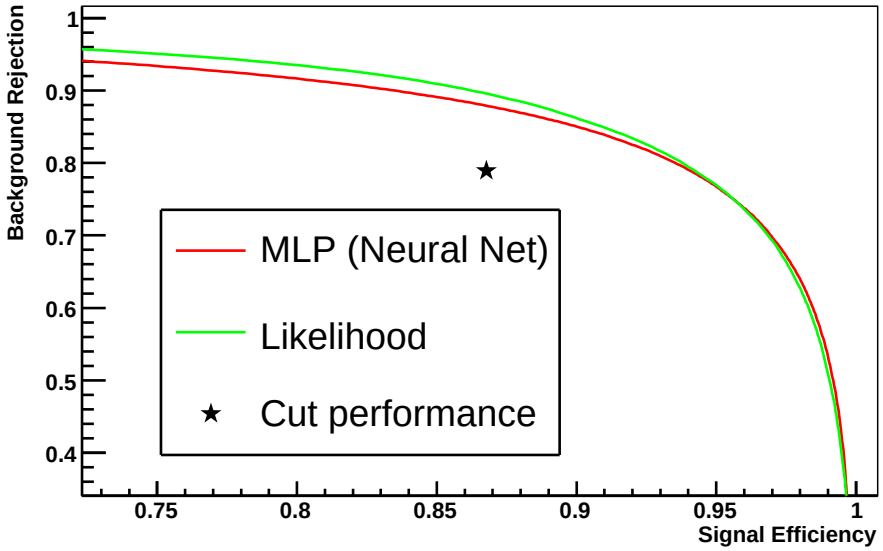

FIG. 9: Signal Efficiency vs. Background Rejection plot to measure performance of final, fully-trained classifiers for plug photons. The efficiency and rejection of the standard photon cuts is shown as a black star for comparison.

ID Method	Signal Efficiency	Background Rejection	Cut Value
Standard Cuts	86.8%	78.9%	All Cuts Applied
Neural Net (MLP)	86.8%	87.9%	0.620
Likelihood	86.8%	89.7%	0.062
Neural Net (MLP)	94.1%	78.9%	0.382
Likelihood	94.2%	78.9%	-0.097

TABLE VI: Comparing the performance of standard cuts and newly-trained multivariate ID methods for plug photons. We assess the potential gains in signal efficiency with background rejection maintained at current (cut-based) levels, and gains in background rejection with signal efficiency similarly maintained. Improved signal efficiency of 9% or background rejection of 14% is acheived for the samples used above. Actual performance depends on signal and background composition and may be optimized at some other point. If optimization isn't feasible, choosing one of the above cuts to improve only signal efficiency or background rejection is recommended.

APPENDIX C: USE IN STNTUPLE

The trained MLP has been added to Stntuple code in CVS. In your Stntuple release, you need to update and build the obj and alg libraries. Something like this should work:

```
cvs co -r dev_243 Stntuple/Stntuple/alg/TStnEmMva.hh
cvs co -r dev_243 Stntuple/alg/TStnEmMva.cc
cvs co -r dev_243 Stntuple/alg/dict/TStnEmMva_linkdef.h
cvs update -r dev_243 Stntuple/Stntuple/obj/TStnPhoton.hh
cvs update -r dev_243 Stntuple/obj/TStnPhoton.cc
cvs update -r dev_243 Stntuple/Stntuple/alg/TStntuple.hh
cvs update -r dev_243 Stntuple/alg/TStntuple.cc
gmake Stntuple._obj USESHLIBS=1
gmake Stntuple._alg USESHLIBS=1
```

You need to initialize the multivariate tool:

```
TStnEvent* event = fHeaderBlock->GetEvent();
TStntuple::FillEmMva(event);
```

Then, within the loop over photon objects:

```
float_t Pho_Mva= pho->EmMva();
```

You can then cut on Pho_Mva.

---

[1] F. Abe et al. (CDF), Nucl. Instr. Meth. **A271**, 387 (1988).

[2] P. T. Lukens (CDF IIb) (2003), FERMILAB-TM-2198.

[3] S. M. Wynne (2007), ph.D. Thesis, University of Liverpool, 2007, FERMILAB-THESIS-2007-17.

[4] C. Lester et al. (CDF) (2007), CDF Note 9033.

[5] K. Bland et al. (CDF) (2010), CDF Note 10038.

[6] A. Hoecker et al. (CERN) (2009), arXiv:physics/0703039.

[7] K. Bland et al. (CDF) (2010), CDF Note 10065.

[8] dataset gq0sqd, inclusive photon production with minbias, 3444250 raw events

[9] dataset q8is01, dijet production with dijet $P_T > 40\ GeV/c$, 77093242 raw events