Machine Learning Techniques in the CMS Search for Higgs Decays to Dimuons

Dimitri Bourilkov[1,2,*], *Darin* Acosta[1], *Pierluigi* Bortignon[1], *Andrew* Brinkerhoff[1], *Andrew* Carnes[1], *Sergei* Gleyzer[1], and *Brendan* Regnery[1]
on behalf of the CMS Collaboration

[1]University of Florida, PO Box 118440, Gainesville, FL 32611, USA
[2]Corresponding author

Abstract. With the accumulation of large collision datasets at a center-of-mass energy of 13 TeV, the LHC experiments can search for rare processes, where the extraction of signal events from the copious Standard Model backgrounds poses an enormous challenge. Multivariate techniques promise to achieve the best sensitivities by isolating events with higher signal-to-background ratios. Using the search for Higgs bosons decaying to two muons in the CMS experiment as an example, we describe the use of Boosted Decision Trees coupled with automated categorization for optimal event classification, bringing an increase in sensitivity equivalent to 50% more data.

1 Physics Motivation

Since the discovery of the Higgs boson in 2012 [1, 2], focus has turned towards precisely measuring its properties. The Higgs couplings to the W and Z gauge bosons, as well as the third generation quarks (top and bottom) and tau leptons, have been observed by CMS and ATLAS, and appear to be consistent with the Standard Model (SM) predictions. Finding Higgs decays to a pair of muons with opposite charge ($\mu\mu$) would provide the first evidence of Higgs couplings to fermions outside the third generation.

This search is made difficult by the minuscule branching fraction of Higgs decays to $\mu\mu$, predicted to be ~0.022% by the SM. In addition, there is a large irreducible background from Drell-Yan $\mu\mu$ production, in addition to top quark or W boson pairs decaying to muons. The Higgs signal has a distinctive dimuon invariant mass peak near 125 GeV, which is only ~4 GeV wide, thanks to the excellent muon momentum measurement in CMS. Meanwhile, the invariant mass spectrum for background events falls smoothly in the search region from 110 to 150 GeV. The CMS detector is described in detail in [3]. The results of the 2016 analysis, using 35.9 fb^{-1} of collision data, were published by the CMS collaboration in [4]. In this paper, we describe in more detail how the analysis was optimized for maximum signal sensitivity, utilizing multivariate and machine learning techniques.

2 Signal and Background

The Higgs boson can be produced by several mechanisms with different cross sections, as shown in figure 1 and table 1. The main source of Higgs bosons is gluon-fusion, followed by

[*]e-mail: dimi@ufl.edu

vector boson fusion (VBF) and associated production with a W or Z boson (VH). While VBF production is much less frequent than gluon-fusion, the presence of additional forward jets distinguishes these events from SM backgrounds, making this channel an important component of the overall search strategy.

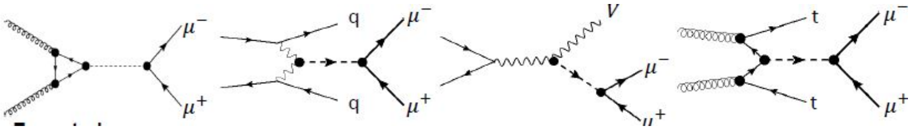

Figure 1. Feynman diagrams for the dominant Higgs production modes: gluon-fusion, Vector Boson Fusion, and associated production with a vector boson (where V = W$^{\pm}$ or Z) or top quark pair ($t\bar{t}$).

Table 1. Higgs cross sections for the different production modes.

	gluon-fusion	VBF	VH	$t\bar{t}H$
Cross sections (pb)	44.08	3.76	2.25	0.51

Drell-Yan decays to muon pairs make up more than 90% of the background, and most closely resemble gluon-fusion signal events, with no additional energetic particles emerging from the collision. The $t\bar{t}$ background and rarer processes with high-momentum jets must be distinguished from VBF signal events.

2.1 Signal Extraction

The amount of Higgs signal in the data is measured by performing a combined signal-plus-background fit to the invariant dimuon mass spectrum in data, using a sum of three Gaussians to model the sharply peaked signal, and a modified Breit-Wigner curve to model the smoothly falling background, as shown in figures 2 and 3. However, with an initial signal-to-background (S/B) ratio of ~0.3% even at the mass point of 125 GeV, an immense amount of data is needed to confirm the presence of signal events. The CMS analysis of 7 and 8 TeV data collected in 2012 [5] divided events into categories based on the transverse momentum (p_T) of the dimuon pair (which is higher for gluon-fusion signal than for Drell-Yan background), or the presence of a high-invariant-mass dijet pair, characteristic of VBF signal events. It also sub-divided the gluon-fusion categories based on the muon pseudorapidity (η), as central muons have better p_T resolution, resulting in a sharper signal mass peak. Performing separate signal-plus-background fits in all of these categories and combining the results significantly increased the search sensitivity relative to a measurement of all candidate dimuon events together.

3 Event Classification

For the analysis of 13 TeV data collected by CMS in 2016, we developed a new event classification based on a larger set of input variables fed into a Boosted Decision Tree (BDT), implemented in the TMVA class [6] of the ROOT analysis package [7]. A binary signal-background separation is computed, yielding a BDT score between -1 and 1, where events close to 1 are more signal-like, and events close to -1 are more background-like.

The signal training set includes the three main production channels: gluon-fusion, VBF, and VH. Variables with some discriminating power for signal-background separation are

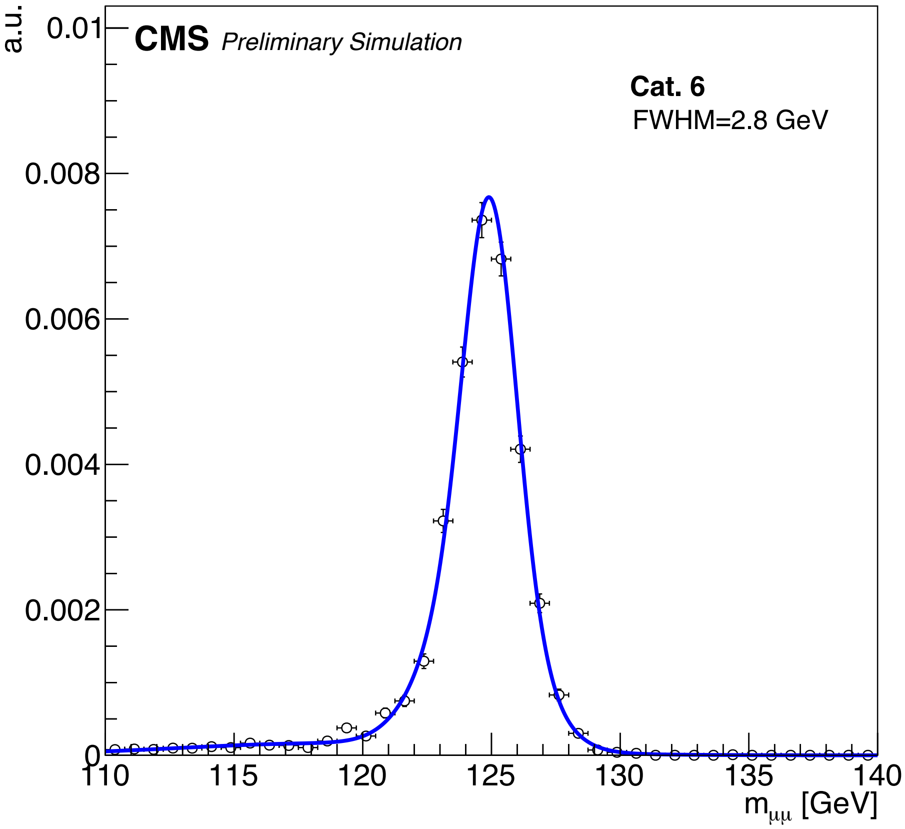

Figure 2. The simulated Higgs dimuon invariant mass fit with a triple Gaussian in event selection categories described in 3.

used, but in all cases the difference between the inclusive signal and background shapes is relatively small. The invariant mass of the dimuon system is excluded, as it will be used independently to measure the amount of signal in each BDT-score-defined category. The kinematic variables selected as input to the BDT are as follows:

- The p_T and η of the dimuon system
- The $|\delta\eta|$ and $|\delta\phi|$ between the muons
- The η values of the two highest–p_T jets
- The masses of the two highest–mass dijet pairs
- The $|\delta\eta|$ between the jets in the two highest–mass pairs
- The number of jets with $|\eta| < 2.4$ and $|\eta| > 2.4$
- The number of jets identified as coming from b-hadrons [8]
- The missing transverse energy E_T.

The dimuon p_T and η are most important, as gluon-fusion signal pairs tend to be more central and have higher p_T than the Drell-Yan background. The dijet separation in η and invariant mass are crucial for clearly identifying the small fraction of VBF signal events.

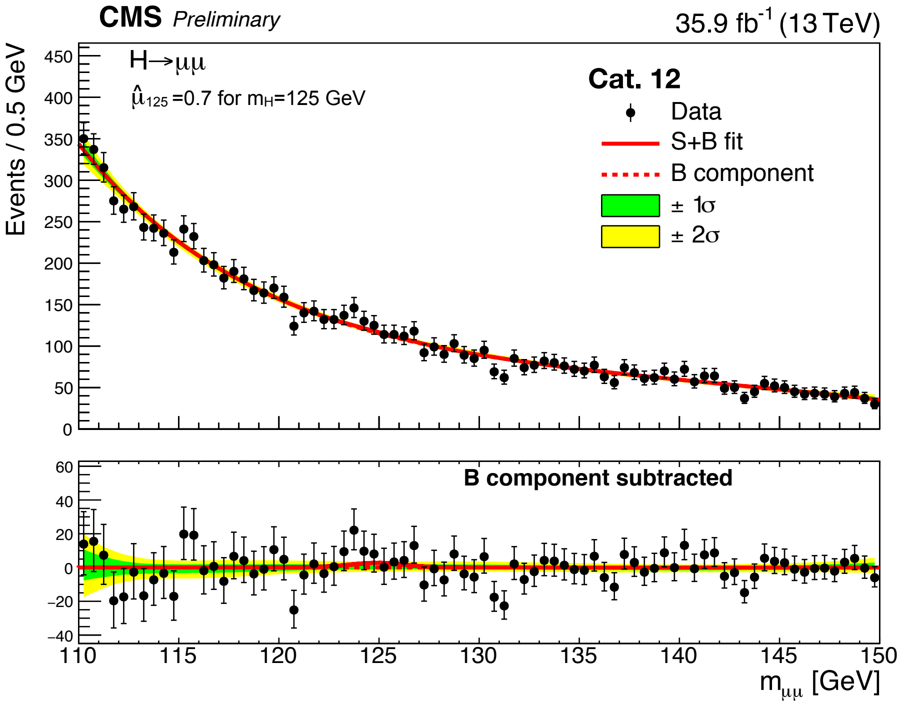

Figure 3. A a combined signal-plus-background fit to the data in event selection categories described in 3.

Events with b-tagged jets and significant missing E_T most likely come from $t\bar{t}$ background events and are assigned a low score.

The training is based on one million simulated events for the various channels, fully reconstructed in the CMS detector. The signal sample is split into three independent sets: one for training, a second for testing, and a third completely independent - to avoid any bias - for the final measurement. The background samples are typically split in 75% for training and 25% for testing. The final measurement does not use simulated background, but rather a direct analytic fit to the data in each category.

The BDT uses 400 trees, gradient boost, and variable splitting at 1000. The receiver operating characteristics (ROC) integral below the curve for signal-background separation is 0.72. The BDT response, transformed in quantiles with a uniform distribution in the sum of expected signal events from all production modes, is shown in figure 4. The VBF signal events have the highest BDT scores, with the best signal-to-background ratio. Gluon fusion events have generally higher scores than the dominant Drell-Yan background, and $t\bar{t}$ events congregate at the lowest end of the spectrum.

Because the final signal measurement is made using a fit to the dimuon invariant mass spectrum, it is important that high BDT scores are not correlated to signal-like mass values. If such a correlation existed, real background events in the highest BDT bins would be biased towards 125 GeV, and could mimic an excess of signal events even if no true signal events were present. To confirm that such a bias does not exist, we evaluate the BDT on simulated signal events generated with Higgs boson masses of 120, 125, and 130 GeV. As the BDT output distribution looks identical for these three samples, we confirm that the BDT has not

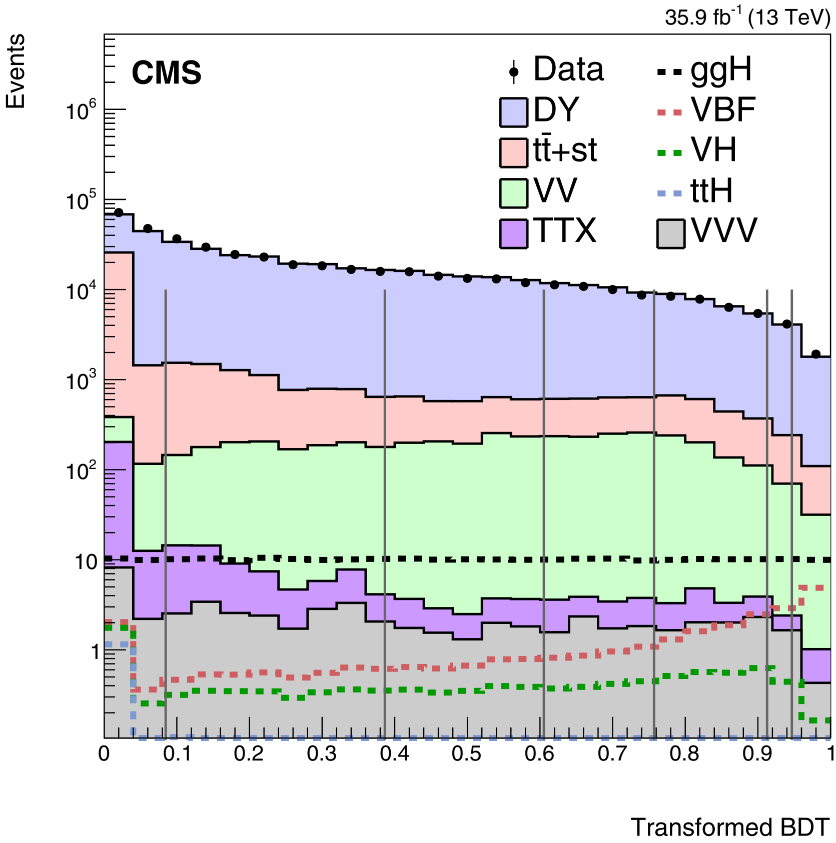

Figure 4. Normalized BDT output, transformed to be uniform for the sum of expected signal events [4].

"learned" the true signal mass, at least to a resolution less than 10 GeV, which is much larger than the 4 GeV dimuon mass resolution at 125 GeV.

4 Decision Tree Auto-Categorizer

Once the BDT scores for each event are obtained, further improvement in the signal-to-background separation is possible by taking into account the fact that more central events in the CMS barrel have a better mass resolution than forward events where at least one of the muons is in one of the two CMS endcaps. The basic idea is to "greedily" optimize the sensitivity by simultaneously categorizing events based on the BDT score (from -1 to 1) and the maximum muon $|\eta|$, which is directly correlated to the mass resolution: 2.8 to 7.6 GeV full width at half-maximum (FWHM) for $|\eta|$ from 0 to 2.4.

To compute the expected measurement sensitivity, the simulated signal (S) and background events (B) are divided in 0.5 GeV bins in $\mu\mu$ mass for the region from 120 to 130 GeV. The expected signal significance from each bin is given by $S/\sqrt{B}$, and the values from each bin are added in quadrature:

$$\text{Net Significance}^2 = \sum_{C,i} S^2_{C,i}/B_{C,i} \qquad (1)$$

where the sum runs over the categories C and the mass bins i.

The automated categorization procedure is performed in steps. We start with one (inclusive) category C0 with fine mass binning, and check all possible binary cut values on muon $|\eta|$ or event BDT score to find the value giving maximum gain for a split C0→C1+C2:

$$\text{Gain} = \sum_i S^2_{C1,i}/B_{C1,i} + \sum_i S^2_{C2,i}/B_{C1,i} - \sum_i S^2_{C0,i}/B_{C0,i} \qquad (2)$$

In following iterations, we repeat the procedure on the new set of categories, "greedily" going for the maximal gains by splitting one category at a time. We stop the procedure when the gain from one additional cut is no longer significant ($\sim$ 1%).

4.1 Final Categories

At the end of the auto-categorizer procedure, a simplification of the cut boundaries in $|\eta|$ and BDT scores is performed by rounding some of the cuts. We have checked that no sizable loss of sensitivity is introduced when using the simplified boundaries shown in figure 5. In this way we arrive at 15 event categories. The relative gain in sensitivity is 23% compared to the simple categories based on $\mu\mu$ p_T, dijet mass, and muon η used in the CMS searches at 7 and 8 TeV, equivalent to a dataset 50% larger than the one actually collected.

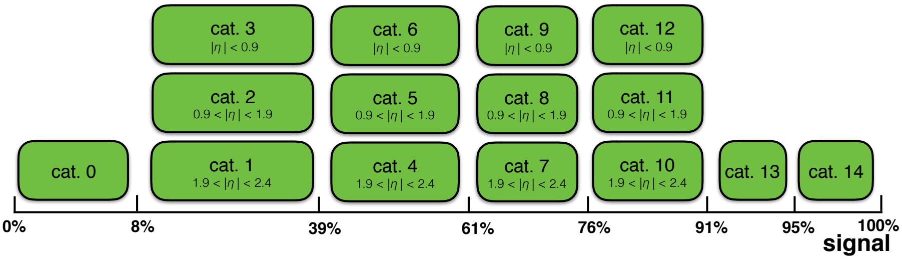

Figure 5. Simplified category table. BDT response quantile in [%] and $|\eta|$ ranges for categories after simplification are shown.

The expected signal and estimated background yields for the final categories, ordered from least to most sensitive, are shown in table 2. The $S/\sqrt{B}$ ratio ranges from 0.12 for category 0 to 0.48 for the most sensitive category 14 with the most signal-like BDT scores.

5 Conclusions and Future Work

The BDT-based categorization helps to enhance the separation power of signal versus background for the difficult search of the Higgs boson decaying to two muons. Using the full data set collected from 2016 to 2018 in Run 2 of the LHC, the experiments will approach the discovery zone for this important decay to the second generation fermions, with an expected significance around 2 standard deviations from the no-signal hypothesis. In order to further increase the sensitivity, deep neural networks with several layers and multi-node architectures are being explored to improve on the already impressive performance of the BDT

Table 2. The optimized event categories, the product of acceptance and selection efficiency in % for the different production processes, the total expected number of SM signal events (m_H = 125 GeV), the estimated number of background events per GeV at 125 GeV, the FWHM of the signal peak, and the $S/\sqrt{B}$ ratio within the FWHM of the expected signal distribution.

BDT response		Muon $\vert\eta\vert$	ggH	VBF	WH	ZH	$t\bar{t}H$	Signal	Bkg./GeV	FWHM	$S/\sqrt{B}$
quantile [%]		range	[%]	[%]	[%]	[%]	[%]		@125 GeV	[GeV]	@ FWHM
0	8	0.0-2.4	4.9	1.3	3.3	6.3	32	21.2	3.13×10^3	4.2	0.12
8	39	1.9-2.4	5.6	1.7	3.9	3.5	1.3	22.3	1.34×10^3	7.2	0.16
8	39	0.9-1.9	10	2.8	6.5	6.4	5.2	41.1	2.24×10^3	4.1	0.29
8	39	0.0-0.9	3.2	0.8	1.9	2.1	3.5	12.7	7.83×10^2	2.9	0.18
39	61	1.9-2.4	2.9	1.7	2.7	2.7	0.3	11.8	4.37×10^2	7.0	0.14
39	61	0.9-1.9	7.2	3.3	6.1	5.2	1.3	29.2	9.70×10^2	4.0	0.31
39	61	0.0-0.9	3.6	1.1	2.6	2.2	0.9	14.5	4.81×10^2	2.8	0.26
61	76	1.9-2.4	1.2	1.5	1.8	1.7	0.2	5.2	1.48×10^2	7.6	0.11
61	76	0.9-1.9	4.8	3.6	4.5	4.4	0.7	20.3	5.12×10^2	4.2	0.29
61	76	0.0-0.9	3.2	1.6	2.3	2.1	0.6	13.1	3.22×10^2	3.0	0.28
76	91	1.9-2.4	1.2	3.1	2.2	2.1	0.2	5.8	1.04×10^2	7.1	0.14
76	91	0.9-1.9	4.4	8.7	6.2	6.0	1.1	20.3	3.60×10^2	4.2	0.35
76	91	0.0-0.9	3.1	4.0	3.8	3.6	0.9	13.7	2.36×10^2	3.2	0.34
91	95	0.0-2.4	1.7	6.4	2.5	2.6	0.5	8.6	96.0	4.0	0.28
95	100	0.0-2.4	2.0	19	1.5	1.4	0.7	13.7	83.4	4.1	0.48
0	100	0.0-2.4	59	61	51	52	49	253	1.30×10^4	3.9	

techniques. Additional promising avenues include the use of extended sets of discriminating variables, especially those targeting the rare VH and $t\bar{t}H$ production modes. Along with the multivariate discriminators, a more sophisticated automated categorization is being considered, accounting explicitly for the fit function uncertainty in the background estimate in each category. Taken together, the combined BDT plus auto-categorization approach serves as a model for how to achieve the maximum sensitivity to some of the rarest events in nature.

References

[1] S. Chatrchyan *et al.* [CMS Collaboration], Phys. Lett. B **716**, 30 (2012) doi:10.1016/j.physletb.2012.08.021 [arXiv:1207.7235 [hep-ex]].

[2] G. Aad *et al.* [ATLAS Collaboration], Phys. Lett. B **716**, 1 (2012) doi:10.1016/j.physletb.2012.08.020 [arXiv:1207.7214 [hep-ex]].

[3] CMS Collaboration, JINST **3**, S08004 (2008).

[4] A. M. Sirunyan *et al.* [CMS Collaboration], Phys. Rev. Lett. **122** no.2, 021801 (2019) doi:10.1103/PhysRevLett.122.021801 [arXiv:1807.06325 [hep-ex]].

[5] V. Khachatryan *et al.* [CMS Collaboration], Phys. Lett. B **744**, 184 (2015) doi:10.1016/j.physletb.2015.03.048 [arXiv:1410.6679 [hep-ex]].

[6] Hoecker, Andreas and Speckmayer, Peter and Stelzer, Joerg and Therhaag, Jan and von Toerne, Eckhard and Voss, Helge, arXiv:physics/0703039, 2007.

[7] Rene Brun and Fons Rademakers, Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A **389**, 81-86 (1997). See also root.cern.ch/.

[8] CMS Collaboration, JINST **13**, P05011 (2018).