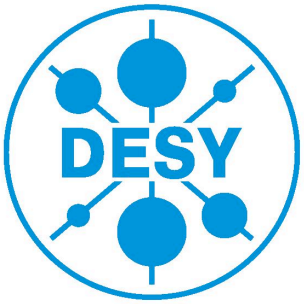

Improvement of event categorisation

for the Higgs produced in $t\bar{t}H$ and decaying into $b\bar{b}$

Supervisors: J. Katzy, J. S. Keller

H. Gan

Department of Physics, University of Heidelberg

September 9, 2015

Abstract

A study to define categories of events which contain the Standard Model Higgs boson produced in association with a pair of top quarks and decaying into a pair of b-quarks, $t\bar{t}H(\rightarrow b\bar{b})$, is presented. The analysis is based on simulated pp collision data at $\sqrt{s} = 13$ TeV corresponding to an integrated luminosity of 5 fb^{-1}. The study focuses on events including exactly one electron or muon. Events are categorised to nine regions depending on the number of jets and b-tags. In order to optimise the sensitivity, mixed-working-points cuts are used to define the number of b-tags in each category. Two regions ($\geq$6j,$\geq$4b) and ($\geq$6j,3b) are found to be signal-rich according to the signal-rich conditions S/B $> 1\%$ and S/$\sqrt{B} > 0.3$. The results show that in region ($\geq$6j,$\geq$4b), the new categories with mixed-WPs cuts give the better S/$\sqrt{B}$ ratios and higher statistics compared to the categories with constant-WP cuts.

Contents

1 Introduction

The Standard Model(SM) Higgs boson was discovered by the ATLAS[1] and CMS[2] collaborations in July 2012. In Run1 of LHC, the constraints on $t\bar{t}H$ have been set.[3] To understand all properties of the new boson, it is important to study the particle in as many production and decay modes as possible. The following decay channels have been measured at a mass of around 125 GeV.

- $H \to \gamma\gamma$
- $H \to ZZ^{(*)} \to 4l$
- $H \to WW^{(*)} \to l\nu l\nu$
- $H \to \tau\tau$
- $H \to b\bar{b}$

In particular, its coupling to heavy quarks is a strong focus of current experimental searches. Studying the SM Higgs boson produced in association with a top-quark pair($t\bar{t}H$) with subsequent Higgs decay into bottom quarks($H \to b\bar{b}$) addresses heavy-quark couplings.
Because of the large mass of the top quark, the Yukawa coupling of the top quark y_t is much stronger than that of the other quarks. Thus the observation of $t\bar{t}H$ would allow the direct measurement of this coupling. Plus, y_t is expected to be close to unity, which is the quantity that might give insight into the scale of new physics. This study

is designed to be sensitive to the $t\bar{t}H(\to b\bar{b})$ decay which is the dominant decay mode for SM Higgs boson with a mass of 125 GeV. The first two diagrams of Fig.1 show two examples of tree-level Feynman diagrams for $t\bar{t}H$ production with a subsequent $H \to b\bar{b}$ decay. The main source of background comes from the production of $t\bar{t}$. The dominant source is $t\bar{t} + b\bar{b}$ production which ends up with the same final-state signature as the signal. An example of the dominant background processes is shown in the last diagram of Fig.1. One can notice that $b\bar{b}$ pair in the final-state is not from Higgs boson, but from gluon radiation.
The second contribution of background arises from $t\bar{t}$ production in association with light-quark(u, d, s) or gluon jets and with c-quarks, respectively referred to as $t\bar{t}$ +

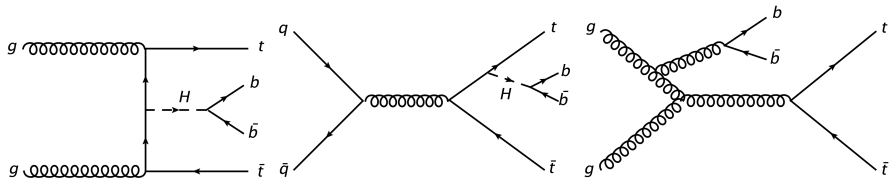

Figure 1: The tree-level Feynman diagrams for the signal and background.[3]

light and $t\bar{t}$ + $c\bar{c}$ background. The size of the second contribution depends on the misidentification rate of the algorithm used to identify b-quark jets.[3]
This report is organised as follows. The event selection is briefly described in Section.2 with the simulated samples used for the analysis. Also, the working points of b-tagging for the study is described. The event categorisation is presented in Section.3 and the mixed-WPs cuts for the improvement of categorisation are introduced. The results of categorisation are reported in Section.4 while Section.5 provides the conclusions.

2 Data samples and event selection

2.1 Data samples

All the following calculations are performed with samples corresponding to 13 TeV proton-proton(pp) collisions and normalised to an integrated luminosity $L = 5fb^{-1}$. The signal($t\bar{t}H$) sample is simulated by aMC@NLO and Herwig++. $t\bar{t}H$ sample uses the CT10 PDF in the ME and CTEQ6L1 in the parton shower with the UEEE5 tune for Herwig++. In $t\bar{t}H$ sample, all Higgs boson decays are allowed, while the top quark pair is filtered to decay semileptonically. The background($t\bar{t}$) sample is simulated by POWHEG and PYTHIA. $t\bar{t}$ sample uses the CTEQ6L1 PDF and the Perugia2012 tune for PYTHIA. $t\bar{t}$ events are filtered to be non-all-hadronic. The details about the samples are summarised in Table1. In addition, the backgrounds from top pair production events associated with a vector boson $t\bar{t}$ + V and non-$t\bar{t}$ are not included in the samples, since they were not available at the time of the study. However, these backgrounds are expected to contribute less than 15% to the total background.

	Signal	Background
Event generator	aMC@NLO + Herwig++	POWHEG + PYTHIA
Sample size	520322	1997974
Filter	inc_semil	non-all-had

Table 1: Details of the samples.

2.2 Event selection

The event selection can be divided to two steps. In the first step, events are selected according to the $t\bar{t}$ production signature. Then in the next step, the selected top pair events are categorised to search for the $t\bar{t}H$ signature.
To select $t\bar{t}$ events, one considers the properties of the top quark. Apart from the large mass, the top quark is also singular because it decays before it hadronises. The top quark decays almost exclusively to a W boson and a b quark, with the fraction determined by the near-unity value of CKM quark mixing matrix element V_{tb} (≈ 0.9992). Subsequently, the W boson decays to a charged lepton and its associated neutrino, or to a quark-antiquark pair($q\bar{q}$)[6]. Thus the final states of $t\bar{t}$ events can be summarised as follows

Classification	Final state
lepton + jets	$l^+\nu bq\bar{q}'b$ or $l^-\bar{\nu}b\bar{q}q'b$
alljets	$\bar{q}q'b\bar{q}q'\bar{b}$
dileptons	$l^+\nu bl^-\bar{\nu}\bar{b}$

Table 2: Decay modes of $t\bar{t}$.[6]

In this study, we focus on the single-lepton channel with the final state corresponding to the lepton + jets category in Table.2. Events in single-lepton channel are required to have exactly one identified electron or muon, which means only one of two W bosons decays leptonically in the final state as shown in Fig.2. Leptons and jets are required to have $|\eta| < 2.5$ and $p_T > 25$ GeV.

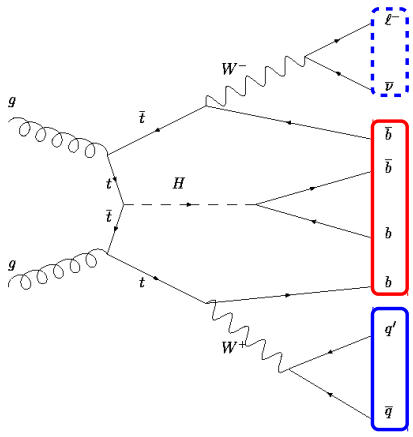

Figure 2: A Feynman diagram example of the final state of the signal process.

2.3 *b*-tagging

Since there are several *b*-jets in the final-state of the signal process(see Fig.2), *b*-tagging is an important issue for the event selection. *b*-tagging refers to the identification of *b*-quark jets. The basic *b*-tagging algorithms use charged particle tracks to produce a set of variables which discriminate between different jet flavours. The algorithms are based on the fact that *b*-hadrons take a long time to decay compared to other hadrons. ATLAS uses three distinct *b*-tagging algorithms, which are as follows :

- Impact parameter of associated tracks based algorithm
- Inclusive secondary vertex reconstruction algorithm
- *b*- $\rightarrow$ *c*- $\rightarrow$ light hadron decay chain multi-vertex reconstruction algorithm

The variables obtained from the three basic algorithms are combined afterward using a boosted decision tree(BDT) algorithm to separate b-jets from light(u, d, s-quark or gluon jets) and c-jets.
The MV2c20 variable is defined as the output of such a BDT with the training performed assigning b-jets as signal and a mixture of 80% light-flavour jets and 20% c-jets as background. [5]
The Fig.3 shows the distribution of MV2c20 variable which is included in our data samples. If the output of MV2c20 variable is close to 1, the corresponding jet is more b-like, and if the output is close to -1, the jet is less b-like.
As shown in Fig3, 60% of b-jets lie in the 60% b-tag efficiency region. Equivalently, it corresponds to working point, WP60, with the MV2c20 cut value of 0.4496.
Four different working points are considered for the event selection and categorisation with different combinations which will be explained in Section3 :

Working point	Cut value
WP60	0.4496
WP70	-0.0436
WP77	-0.4434
WP85	-0.7787

Table 3: Working points and corresponding cut values

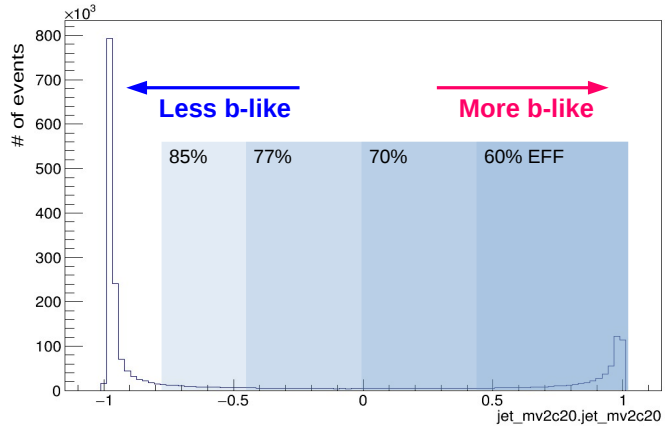

Figure 3: The distribution of MV2c20 variable.

3 Event categorisation

The selection for the $t\bar{t}$ events is described in the previous section. Afterward, selected events are categorised to search for the $t\bar{t}H$ signal as described below. The $t\bar{t}H$ events

have two more b-jets than $t\bar{t}$ events. Therefore, we categorise the events based on the number of jets and b-jets into regions with mixed-WPs cuts.

3.1 Regions

Selected events are classified into exclusive categories, depending on the number of reconstructed jets and b-quark jets(b-tagged jets) which are identified by the b-tagging algorithm. For b-tagging, we use the MV2c20 variable in this study, which already has been discussed in Section.2. By definition, a given region with n jets and m b-jets is denoted as (mj,nb). The regions with a signal-to-background ratio $S/B > 1\%$, where S and B denote the number of expected signal and background, and $S/\sqrt{B} > 0.3$ are referred to as 'signal-rich-regions', as they provide most of the sensitivity to the signal. We define nine regions which are consistent with the previous study[3] as follows:

$$\begin{array}{lll} \text{(4j,2b)} & \text{(4j,3b)} & \text{(4j,4b)} \\ \text{(5j,2b)} & \text{(5j,3b)} & \text{(5j,}\geq\text{4b)} \\ \text{(}\geq\text{6j,2b)} & \text{(}\geq\text{6j,3b)} & \text{(}\geq\text{6j,}\geq\text{4b)} \end{array} \tag{1}$$

After the categorisation, neural networks(NN) are employed in the regions to separate the signal from the background. Thus our final goal of event categorisation is to find the best way to define signal-rich-regions, i.e. the regions with high $S/\sqrt{B}$ and high statistics to improve the sensitivity of the NN.

3.2 Mixed-WPs cuts

Traditionally, b-jets are identified with a constant-WP(one WP). In 8 TeV search[3], b-jets are identified with the MV1 variable at WP70. In this study, we suggest a categorisation based on non-constant cuts on the b-tagging discriminant of the various jets to optimise the sensitivity. The number of b-jets in each category is defined with mixed-WPs cuts. There are some early studies of mixed-WPs cuts which are performed on 4b-tagged regions.[7][8]
The basic idea of mixed-WPs cut is as follows. Consider a constant-WP cut, for instance, 3b-tags pass over WP70 as shown in Fig.4 left. However, if you consider one more WP, WP60, at the same time, then further classification is possible as shown in Fig.4 right. As a result, the constant-WP cut (3b@WP70) can be classified into four possible 'combinations 'with 2WPs cuts.

- Combination 1: 3 b-tags lie between WP70∼WP60
- Combination 2: 2 b-tags lie between WP70∼WP60 and 1b-tag passes WP60
- Combination 3: 1 b-tags lie between WP70∼WP60 and 2b-tag pass WP60
- Combination 4: 3 b-tags pass WP60

With four combinations of b-tags, we can perform a new event categorisation as shown in Fig.5. Combinations 2, 3 and 4 can be classified to a new category (3b@WP77,

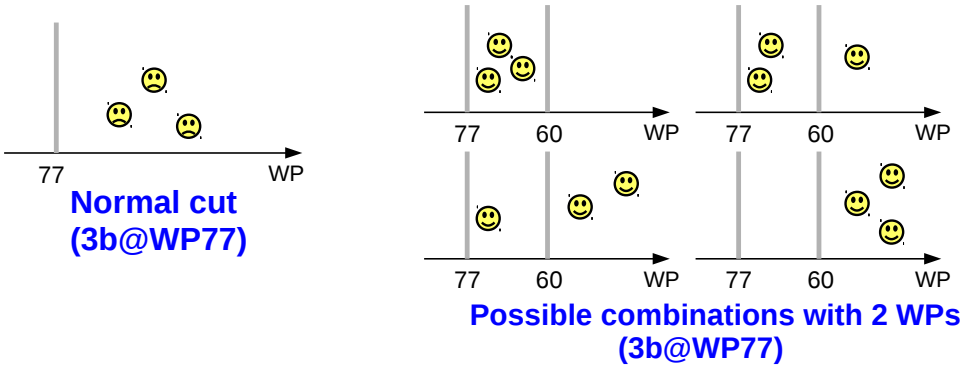

Figure 4: Application of a constant-WP cut and possible combinations for categorisation of three b-tags with 2WPs.

$\geq$1b@WP60) and Combinations 3 and 4 can be classified to another category (3b@WP77, $\geq$2b@WP60).

Figure 5: Two examples of categories with mixed-WPs cuts.

For each category, we use 2WPs to define the number of b-jets. The WP with higher b-tagging efficiency is called the 'loose cut' while the WP with the lower efficiency is called the 'tight cut'. With four b-tag working points mentioned in Table.3, we have made different combinations of mixed-WPs cuts. They are listed in Table.4

2b-tagged(6)	3b-tagged(12)	
(≥1b@WP60,2b@WP70)	(≥1b@WP60,3b@WP70)	(≥2b@WP60,3b@WP70)
(≥1b@WP60,2b@WP77)	(≥1b@WP60,3b@WP77)	(≥2b@WP60,3b@WP77)
(≥1b@WP60,2b@WP77)	(≥1b@WP60,3b@WP77)	(≥2b@WP60,3b@WP77)
(≥1b@WP70,2b@WP77)	(≥1b@WP70,3b@WP77)	(≥2b@WP70,3b@WP77)
(≥1b@WP70,2b@WP77)	(≥1b@WP70,3b@WP77)	(≥2b@WP70,3b@WP77)
(≥1b@WP77,2b@WP77)	(≥1b@WP77,3b@WP77)	(≥2b@WP77,3b@WP77)

4b-tagged(18)		
(≥1b@WP60,≥4b@WP70)	(≥2b@WP60,≥4b@WP70)	(≥2b@WP60,≥4b@WP70)
(≥1b@WP60,≥4b@WP77)	(≥2b@WP60,≥4b@WP77)	(≥2b@WP60,≥4b@WP77)
(≥1b@WP60,≥4b@WP77)	(≥2b@WP60,≥4b@WP77)	(≥2b@WP60,≥4b@WP77)
(≥1b@WP70,≥4b@WP77)	(≥2b@WP70,≥4b@WP77)	(≥2b@WP70,≥4b@WP77)
(≥1b@WP70,≥4b@WP77)	(≥2b@WP70,≥4b@WP77)	(≥2b@WP70,≥4b@WP77)
(≥1b@WP77,≥4b@WP77)	(≥2b@WP77,≥4b@WP77)	(≥2b@WP77,≥4b@WP77)

Table 4: Possible combinations for 2b-, 3b- and 4b-events

If we take constant-WP cuts into account, there are 144 possible combinations in total. In order to find out possible signal-rich-regions, we proceed the calculation of S/B and S/$\sqrt{B}$ ratio for each combination.

4 Results

To calculate S/B and S/$\sqrt{B}$, the first step is to calculate expected signal S and background B respectively with the proper normalisation. In the study, S and B are normalised with an integrated luminosity L = 5 fb^{-1}.

$$S = L * \sigma_{t\bar{t}H} * BR(t\bar{t}H) * \frac{n_{t\bar{t}H}}{N_{t\bar{t}H}} \tag{2}$$

$$B = L * \sigma_{t\bar{t}} * BR(t\bar{t}) * \frac{n_{t\bar{t}}}{N_{t\bar{t}}} \tag{3}$$

where σ, BR respectively indicates cross section and branching ratio. n is the raw number of events in each category while N is the sum of event weights representing the total number of events in the sample. Some constants used to do the S/B and S/$\sqrt{B}$ ratio calculations are written in Table.5.

	Signal($t\bar{t}H$)	Background($t\bar{t}$)
Cross section σ	0.4467 pb	831.76 pb
Branching ratio BR	43.9%	54.3%
The sum of event weights N	520322	1997974

Table 5: Cross section, branching ratio and the sum of events weights for signal and background.

4.1 Signal-to-background ratio

S/B and S/$\sqrt{B}$ are calculated respectively for regions with 2b-, 3b- and 4b-tagged jets as shown in Fig.6, Fig.7 and Fig.8. Each line in individual plot indicates the results of a region. For example, for 2b-tagged events plot, the green line shows the result of region (4j,2b).
If we apply the signal-rich-conditions (S/B > 1% and S/$\sqrt{B}$ > 0.3), there are only two signal-rich-regions: (≥6j,≥4b) and (≥6j,3b) and 22 signal-rich-categories out of 144 possible categories. The region (5j,≥4b) is also promising, since it passes S/B condition and the value of S/$\sqrt{B}$ is close to 0.3.

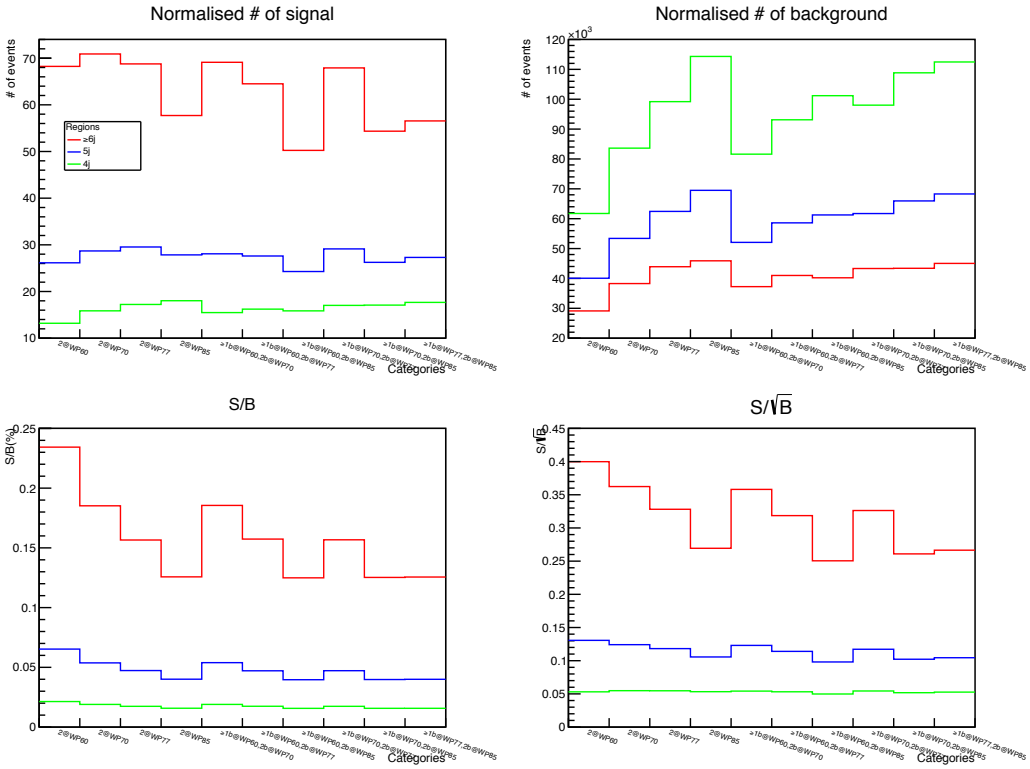

Figure 6: Normalised number of signal and background, calculated S/B and S/$\sqrt{B}$ for 2b-events.

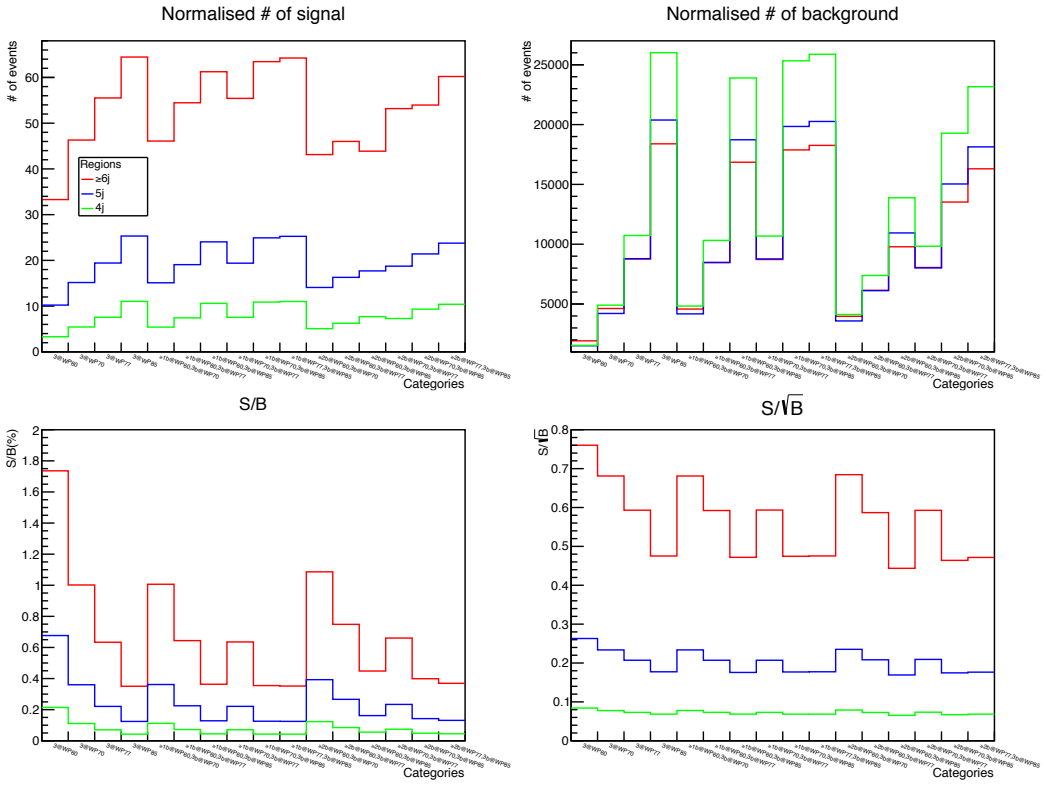

Figure 7: Normalised number of signal and background, calculated S/B and S/$\sqrt{B}$ for $3b$-events.

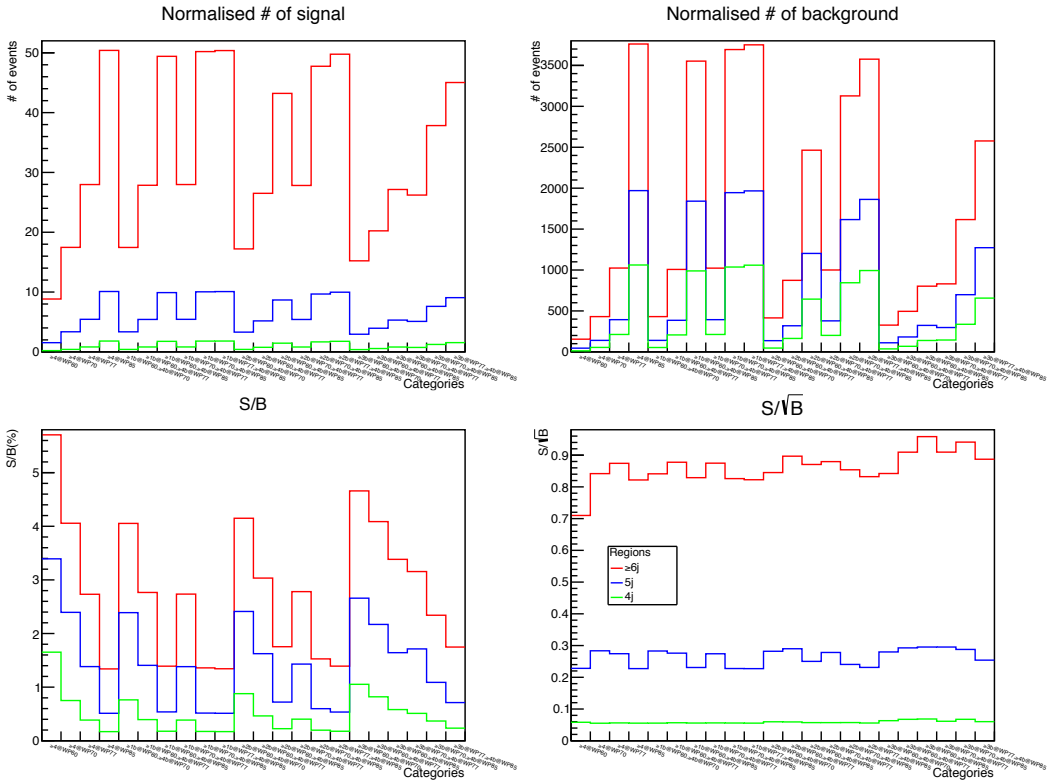

Figure 8: Normalised number of signal and background, calculated S/B and S/$\sqrt{B}$ for 4b-events.

For the normalised number of background plot, we can clearly observe that the number of events(y-axis) is going up for first four bins. This is exactly what we expect to see, since the b-tagging cut is getting looser for the first four bins. Also, there is obvious tendency that S/B ratio is getting lower if we apply looser cuts.

4.2 Scatter plots

On the top of the study in 4.1, now we know there are two promising signal-rich-regions: (≥6j,≥4b) and (≥6j,3b) with 22 different possible combinations. Amongst 22 possible combinations, we want to find the best combination with high S/$\sqrt{B}$ and high statistics for the subsequent NN. The large number of events allows us to use finer binning which helps improve the sensitivity of the NN. To achieve this goal, we draw scatter plots with S/$\sqrt{B}$ on x-axis and the number of events (the sum of S and B) on y-axis.
If there are several possible combinations with about the same values of S/$\sqrt{B}$ ratio,

the best combination can be one with the highest expected number of events. For instant, in Fig.9(upper) there are a bunch of combinations with $S/\sqrt{B}$ around 0.47, and we see that the combination (3b@WP85) is the best among them and one can directly veto the combinations below (3b@WP85) since they have fewer events. Similarly, if there are several combinations with the same amount of expected number of events, the best combination should be the one with highest $S/\sqrt{B}$ ratio.

4.3 Improvement

Comparing the results of constant-WP cuts with that of mixed-WPs cuts, in region ($\geq$6j,3b), there is no real gain. The combinations with the highest expected number of events and $S/\sqrt{B}$ don't show much discrimination between constant-WP cuts and mixed-WPs cuts. As shown in Fig.9, the data points with constant-WP cuts and mixed-WPs cuts are really close, which means they have about the same amount of expected number of events and $S/\sqrt{B}$ values. However, for the region ($\geq$6j,$\geq$4b), there is some improvement. Especially comparing the constant-WP cut (4b@WP77) to the mixed-WPs cut ($\geq$3b@WP77,$\geq$4b@WP85), the latter gives much better results both in terms of $S/\sqrt{B}$ ratio and the expected number of events as shown in Table.6.

	(4b@WP77)	($\geq$3b@WP77,$\geq$4b@WP85)
S	28	45
B	1024	2577
$S/\sqrt{B}$	0.87	0.89

Table 6: The most significant improvement of event categorisation in region ($\geq$6j,$\geq$4b).

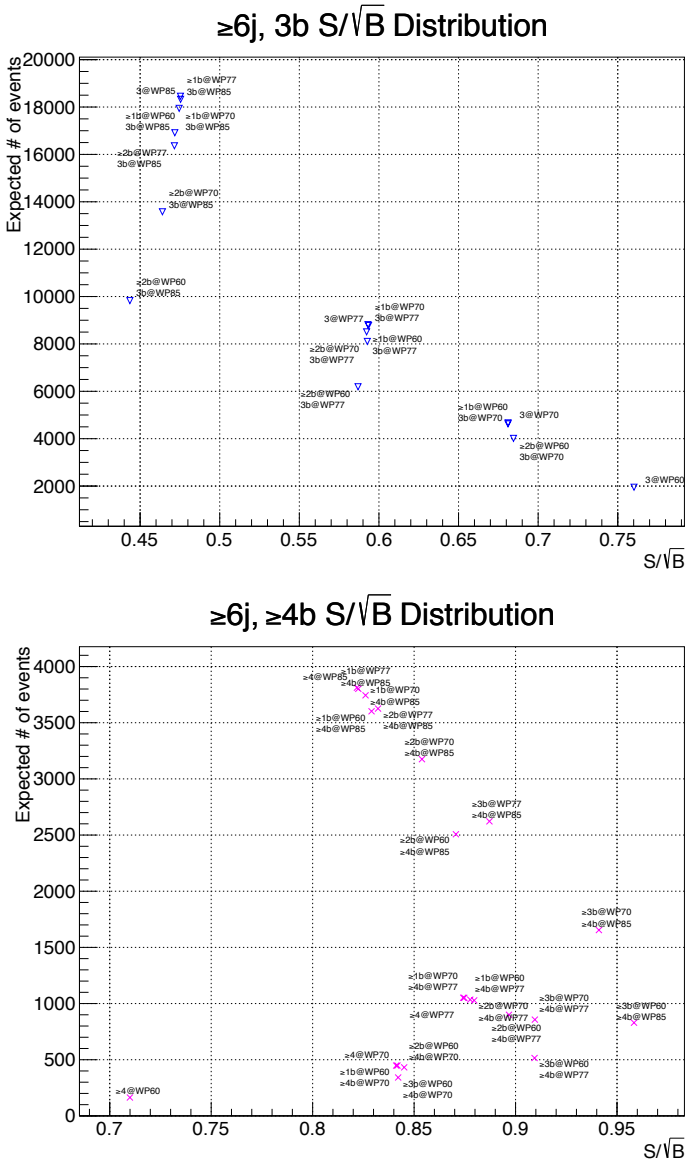

Figure 9: Distribution of $S/\sqrt{B}$ for different combinations in region (≥6j,3b) and (≥6j,≥4b).

Combinations	S	B	S/$\sqrt{B}$
(≥6j,3b)			
(3b@WP60)	33	1918	0.76
(3b@WP70)	46	4623	0.68
(≥1b@WP60,≥3b@WP70)	46	4579	0.68
(≥2b@WP60,≥3b@WP70)	43	3971	0.68
(≥6j,≥4b)			
(≥4b@WP60)	9	155	0.71
(≥4b@WP70)	17	431	0.84
(≥4b@WP77)	28	1024	0.87
(≥4b@WP85)	50	3762	0.82
(≥1b@WP60,≥4b@WP70)	17	431	0.84
(≥1b@WP60,≥4b@WP77)	28	1007	0.88
(≥1b@WP60,≥4b@WP85)	49	3552	0.83
(≥1b@WP70,≥4b@WP77)	28	1023	0.87
(≥1b@WP70,≥4b@WP85)	50	3694	0.83
(≥1b@WP77,≥4b@WP85)	50	3751	0.82
(≥2b@WP60,≥4b@WP70)	17	415	0.85
(≥2b@WP60,≥4b@WP77)	27	874	0.90
(≥2b@WP60,≥4b@WP85)	43	2464	0.87
(≥2b@WP70,≥4b@WP77)	28	1000	0.88
(≥2b@WP70,≥4b@WP85)	48	3127	0.85
(≥2b@WP77,≥4b@WP85)	50	3576	0.83
(≥3b@WP60,≥4b@WP70)	15	326	0.84
(≥3b@WP60,≥4b@WP77)	20	495	0.91
(≥3b@WP60,≥4b@WP85)	27	802	0.96
(≥3b@WP70,≥4b@WP77)	26	831	0.91
(≥3b@WP70,≥4b@WP85)	38	1616	0.94
(≥3b@WP77,≥4b@WP85)	45	2577	0.89

Table 7: The expected signal S, expected background B and S/$\sqrt{B}$ of all signal-rich-categories.

5 Conclusions

The event categorisation for SM Higgs boson produced with a top pair $t\bar{t}H$ is performed with the simulated data corresponding to the pp collision at $\sqrt{s} = 13$ TeV and an integrated luminosity of 5 fb^{-1}. Depending on the number of jets and b-tags, events are categorised into nine different regions.

To optimise the sensitivity, two different WPs are used to define the number of b-tags in each category. Four working points used in the study are : 60%, 70%, 77% and 85%. After the categorisation, S/$\sqrt{B}$ is calculated and there are two signal-rich-regions, (≥6j,≥4b) and (≥6j,3b), which pass the signal-rich-conditions S/B > 1% and S/$\sqrt{B} > 0.3$.

The improvement of categorisation is not obvious in the region (≥6j,3b). However, in the region (≥6j,≥4b), the mixed-WPs cut (≥3b@77%,≥4b@85%) explicitly provides a better possibility of NN than a constant-WP cut (4b@77%) since the new cut gives high statistics and the slightly better $S/\sqrt{B}$ ratio.
For the further studies, we can consider

- To check the flavour composition in categories.
- To check for the possible combinations of non-overlapping categories.
- To use the b-tagging scores as a variable in the NN rather than for categorisation.
- To check the effect of b-tagging uncertainties.

References

[1] 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 (2012) 1, arXiv:1207.7214, 2012

[2] CMS collaboration, *Obsevation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B716 (2012) 30, arXiv:1207.7235, 2013

[3] ATLAS collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks in pp collisions at* $\sqrt{s} = 8TeV$ *with the ATLAS detector at the LHC*, arXiv:1405.1994, 2014

[4] J. Adelman et al., *Search for the Standard Model Higgs boson produced in association with top quarks in pp collisions at* $\sqrt{s} = 8TeV$ *with the ATLAS detector at the LHC*, ATL-COM-PHYS-2013-1659, 2014

[5] ATLAS collaboration, *Expected performance of the ATLAS b-tagging algorithms in Run-2*, ATL-PHYS-PUB-2015-022, 2015

[6] T. Altonen et al., *Combination of the top-quark mass measurements from the Tevatron collider*, arXiv:1207.1069v4

[7] S. Honda et al., *Recovery of 6j2b category*

[8] J. Jovicevie et al., *Jet flavour composition for different b-tagging working point in ttH(bb) semileptonic channel*