PROCEEDINGS OF SCIENCE

Application of Multivariate Analysis in Separation of Higgs Boson Signal at Future e^+e^- Colliders

Ivana Vidaković[a], **Mirko Radulović**[b]* **Jasna Stevanović**[b], **Goran Kačarević**[a]

[a] *"VINCA" Institute of Nuclear Sciences - National Institute of the Republic of Serbia, University of Belgrade, Mike Petrovica Alasa, Belgrade, Serbia*

[b] *Faculty of Science, University of Kragujevac, Radoja Domanovica 12, Kragujevac, Serbia*

E-mail: ivana.vidakovic@vin.bg.ac.rs, mirko.radulovic@pmf.kg.ac.rs

Abstract. Even though the environment at future e^+e^- colliders is practically QCD background free, there is a large number of processes with high cross-sections and/or similar topology as the Higgs signal of interest. Maximization of the achievable precision of measurements in the Higgs sector and beyond calls for optimized event selection with respect to the statistical significance. This is where the Multivariate Analysis (MVA) is employed, separating the signal from numerous backgrounds on the basis of their kinematic and other properties. In this paper, we discuss the basics of MVA, its application and performance, in examples of several Higgs analyses done in our group using full simulation of the CLIC data.

11th International Conference of the Balkan Physical Union (BPU11),
28 August - 1 September 2022
Belgrade, Serbia

*Speaker

1. Introduction

The Standard Model of particle physics is a very successful theory which describes elementary particles and their mutual interactions, leaving however numerous open questions like effective weakness of gravity, nature of the dark matter, neutrino masses, stabilization of scalar masses, unification of forces, baryon asymmetry of the Universe and many more.

High-energy operation (≥ TeV) of future Higgs factories as well as a clean enviroment of e^+e^- collisions enable superior sensitivity to probe various Beyond the Standard Model (BSM) realizations in the Higgs sector and beyond. Future Higgs factories are thus primarily e^+e^- colliders, circular or linear, with available center-of-mass energies up to 3 TeV. The overview of main parameters of future Higgs factories is given in Table 1 [1].

Table 1. Future high-energy e^+e^- colliders parameters.

Collider	Centre-of-mass energy	Integrated luminosity
ILC (International Linear Collider)	250 GeV 500 GeV 1 TeV	2 ab^{-1} 4 ab^{-1} 5 ab^{-1}
CLIC (Compact Linear Collider)	350 (380) GeV 1.4 (1.5) TeV 3 TeV	1 ab^{-1} 2.5 ab^{-1} 5 ab^{-1}
FCC-ee (Future Circular Collider)	240 GeV 350 GeV	5 ab^{-1} 1.5 ab^{-1}
CEPC (Circular Electron Positron Collider)	240 GeV	5.6 ab^{-1}

As will be illustrated, the Higgs signal, in particular in exclusive production and decay modes, is a rare event in comparison to numerous background processes. Despite the ‘clean’ (almost QCD background free) environment at future e^+e^- colliders there is a large number of processes with high cross-sections and similar topologies, even in the most abundant Higgs decay channels. The above calls for a sophisticated optimization of event selection based on simultaneous analysis of multiple observables - multivariate analysis (MVA).

2. Higgs boson as a signal

The cross-sections of main Higgs production processes over the wide span of center-of-mass energies at future Higgs factories are shown in Figure 1 [2]. Some of them have cross-sections as large as hundreds of fb, like Higgsstrahlung at around 200 GeV or WW-fusion above 500 GeV. However, even by choosing Higgs boson decay mode in the most abundant $b\bar{b}$ or W^+W^- channels, numerous multi-jet final state background overwhelms the signal. Some channels, like $H \to \gamma\gamma$ or $H \to \mu^+\mu^-$ have simple experimental signatures in comparison to multi-jet final states, making them distinctive from background, yet they are highly suppressed due to small couplings to Higgs boson occurring in the case of $H \to \gamma\gamma$ decay at the loop level. Both situations call for application of MVA in order to distinct signal from background in statistically most optimal way. Branching ratios (BR) of the SM Higgs boson for several representative decays under studies at

e^+e^- colliders are summarised in Table 2[1]. Among others, these processes are challeenging to MVA separation of signal and background which in case of Higgs to EW bosons decays includes near-by masses of Higgs W and Z-boson jets, while Higgs decays to muons and photons are particularly challenging due to the fact that signal is rare. Leading Feynman diagrams for the processes under considerations are illustrated in Figure 2 [2]. Expected signal and background statistics with the integrated luminosities foreseen at CLIC at 3 TeV and 1.4 TeV center-of-mass energies, is illustrated in Tables 3, 4 and 5. As can be seen, even in the abundant Higgs decay channel like $H \rightarrow ZZ^*$, background dominates over the signal for several orders of magnitude.

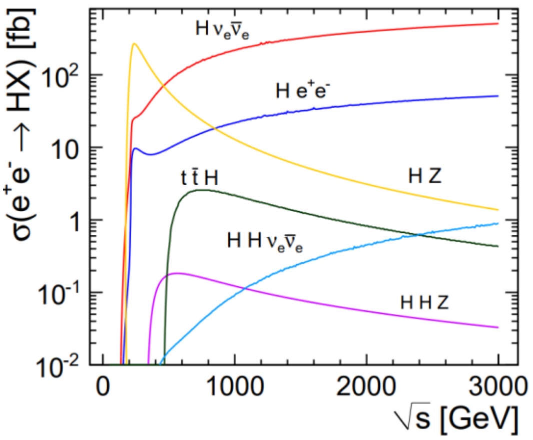

Figure 1. Higgs production cross-sections at different centre-of-mass energies at an e^+e^- collider.

Table 2. The branching ratios for the SM Higgs boson decays.

Decay channel	Branching ratio [%]
$H \rightarrow W^+W^-$	23.1
$H \rightarrow ZZ^*$	2.9
$H \rightarrow \gamma\gamma$	0.23
$H \rightarrow \mu^+\mu^-$	0.021

[1] All processes from Table 2 were studied by the group of authors in the Vinca Institute, including in several studies the authors of this paper, in order to estimate the CLIC sensitivity to measure Higgs BRs in these decay channels.

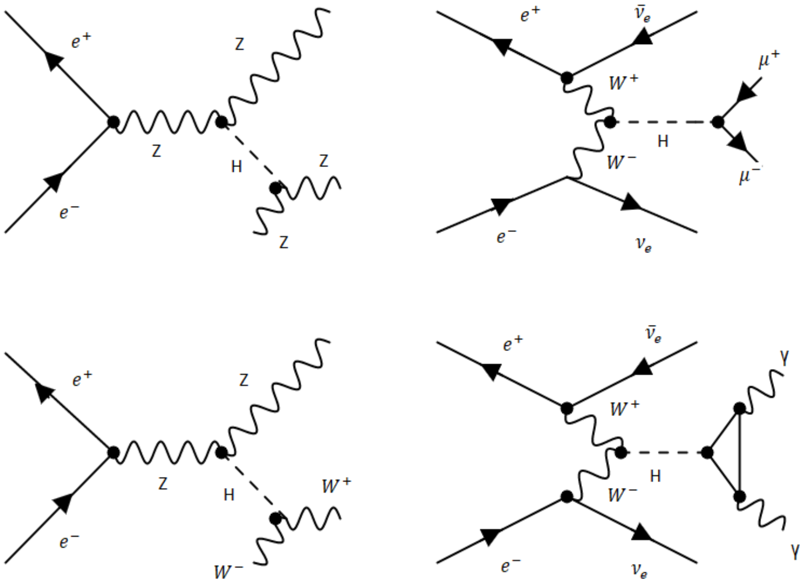

Figure 2. Feynman diagrams for the processes of interest.

Table 3. Signal $H \to ZZ^*$ and background processes with the corresponding cross-sections (σ) and expected number of events (N) at 3 TeV centre-of-mass energy.

Energy	L_{int}	N@5 ab^{-1}	σ (fb)
3 TeV	5 ab^{-1}		
Signal			
$e^+e^- \to H\nu\bar{\nu}$; $H \to ZZ^*$; $ZZ^* \to q\bar{q}l^+l^-$ ($l = e, \mu$)		5650	1.13
Background processes			
$e^+e^- \to H\nu\bar{\nu}$; $H \to WW, WW \to 4q$		$2.15 \cdot 10^5$	43
$e^+e^- \to H\nu\bar{\nu}$; $H \to b\bar{b}$		$1.16 \cdot 10^6$	232
$e^+e^- \to H\nu\bar{\nu}$; $H \to c\bar{c}$		$5.85 \cdot 10^4$	11.7
$e^+e^- \to H\nu\bar{\nu}$; $H \to gg$		$1.75 \cdot 10^5$	35.2
$e^+e^- \to H\nu\bar{\nu}$; $H \to$ others		$4.55 \cdot 10^5$	91
$e^+e^- \to q\bar{q}l^+l^-$		$1.66 \cdot 10^7$	3320
$e^+e^- \to qql\nu$		$2.78 \cdot 10^6$	556
$e^+e^- \to q\bar{q}\nu\bar{\nu}$		$6.60 \cdot 10^6$	1320
$\gamma\gamma \to q\bar{q}l^+l^-$		$1.36 \cdot 10^8$	27200
$\gamma\gamma \to q\bar{q}$		$5.17 \cdot 10^8$	103400
$e^\pm\gamma \to q\bar{q}e$		$6.03 \cdot 10^7$	12060
$e^\pm\gamma \to qq\nu$		$1.38 \cdot 10^8$	27600
$e^+e^- \to q\bar{q}l^+l^-\nu\bar{\nu}$		$1.70 \cdot 10^4$	3.4

Table 4. Signal $H \to \mu^+\mu^-$ and background processes with the corresponding effective cross-sections[2] (σ) and expected number of events (N) at 1.4 TeV centre-of-mass energy.

Energy	$\mathcal{L}_{int}$	N@1.5 ab^{-1}	σ (fb)
1.4 TeV	1.5 ab^{-1}		
Signal			
$e^+e^- \to H\nu\bar{\nu};\ H \to \mu^+\mu^-$		79	$53 \cdot 10^{-3}$
Background processes			
$e^+e^- \to \nu_e\bar{\nu}_e\mu^+\mu^-$		$194 \cdot 10^3$	129
$e^+e^- \to e^+e^-\mu^+\mu^-$		$34.3 \cdot 10^3$	24.5*
$e^\pm\gamma \to e^\pm\mu^+\mu^-$		$1647 \cdot 10^3$	1098*
$e^\pm\gamma \to e^\pm\nu_\mu\bar{\nu}_\mu\mu^+\mu^-$		$45 \cdot 10^3$	30
$\gamma\gamma \to \nu_\mu\bar{\nu}_\mu\mu^+\mu^-$		$243 \cdot 10^3$	162
$e^+e^- \to e^+e^-\nu_\mu\bar{\nu}_\mu\mu^+\mu^-$		$2.4 \cdot 10^3$	1.6

Table 5. Signal $H \to \gamma\gamma$ and background events with the effective cross-sections[3] (σ) and expected number of events (N) at 3 TeV center-of-mass energy.

Energy	$\mathcal{L}_{int}$	N@5 ab^{-1}	σ (fb)
3 TeV	5 ab^{-1}		
Signal			
$e^+e^- \to H\nu\bar{\nu}, H \to \gamma\gamma$		4750	0.95
Background processes			
$e^+e^- \to \gamma\gamma$		$76 \cdot 10^3$	15.2
$e^+e^- \to e^+e^-\gamma$		$1675 \cdot 10^3$	335
$e^+e^- \to e^+e^-\gamma\gamma$		$165 \cdot 10^3$	33
$e^+e^- \to \nu\bar{\nu}\gamma$		$65 \cdot 10^3$	13
$e^+e^- \to \nu\bar{\nu}\gamma\gamma$		$130 \cdot 10^3$	26
$e^+e^- \to q\bar{q}\gamma$		$1050 \cdot 10^3$	210
$e^+e^- \to q\bar{q}\gamma\gamma$		$235 \cdot 10^3$	47

3. Why do we need MVA and how does it work?

Dominance of background over signal, even for the simple final states to reconstruct (for example without multiple jets) calls for a non-trivial exploitation of topological and kinematic properties of an event. Optimized supperssion of background is challenging for the projected precision of measurements in the Higgs sector. Let us show the example of Higgs to di-photon decay illustrated in Table 5.

Events are preselected to define di-photon events as the signal candidate and partially suppress the high-cross section backgrounds such as: $e^+e^- \to e^+e^-\gamma$ and $e^+e^- \to q\bar{q}\gamma$. Only events with exactly two isolated photons with transverse momenta (p_T) greater than 15 GeV are selected. The requirement that both photons have p_T above 15 GeV removes to a great extent

[2] The cross-sections that we noted with (*) are effective in a sense that following conditions are applied on di-muon invariant mass and muon polar angle: 100 GeV < $m_{\mu\mu}$ < 150 GeV, and 8° < θ_μ < 172°.

[3] The cross-sections are effective in a sense that condition 100 GeV < $m_{\gamma\gamma}$ < 150 GeV is applied to any di-photon system found in the barrel region.

reconstructed photons in a signal event that do not originate from the Higgs decay, but rather from a beam radiation (Beamstrahlung). More details about the analysis are given in reference [3]. Preselected signal and background events are further separated using MVA method in TMVA package [4] based on the Gradient Boosted Decision Trees (BDTG). MVA selection is optimizated with respect to the statistical significance defined as:

$$S = \frac{N_S}{\sqrt{N_S + N_B}}$$

where N_S and N_B are number of selected signal and background events. Event will be selected if the MVA output variable of this event is larger than some cut-off value that corresponds to the maximal statistical significance S as illustrated in Figure 3 [3]. In Figure 4 is illustrated signal to background separation with such selection, where the selected signal will be contaminated with some amount of irreducible background impacting the purity of selected signal, while some of the signal will be lost, quantified by selection efficiency. Signal purity and efficiency, background rejection and statistical significance for the training phase on H → γγ signal and background processes from Table 5, are illustrated in Figure 3.

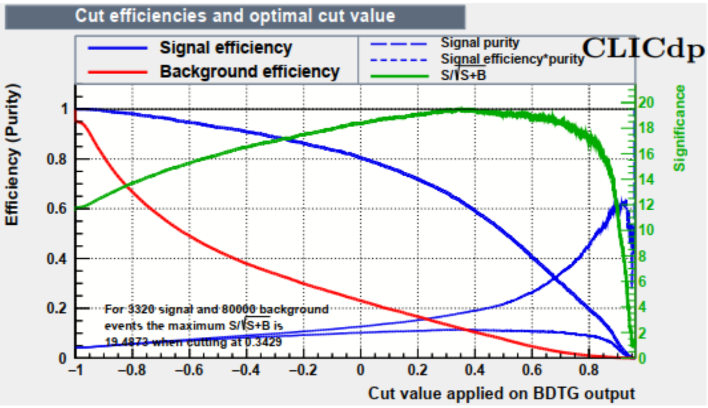

Figure 3. BDTG performance in the training phase to separate H → γγ signal from backgrounds given in Table 5. It can be seen that significance is maximal for the BDTG output cut-off value of ~ 0.34.

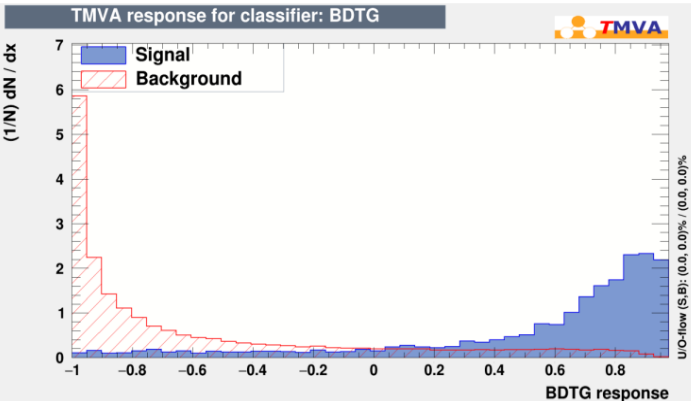

Figure 4. BDTG output variable separating the signal from background, with a cut-off value of ~ 0.34 optimized with respect to the statistical significance.

MVA data analysis is performed in two steps:

- Training phase: the method is learning to separate signal from background on the basis of sensitive observables obtained on Monte Carlo samples for signal and background. Figure 3 illustrates signal purity and efficiency, background rejection rate, statistical significance and BDTG cut-off value;

- Application phase: applying learnt algorithms for separation of signal from background on experimentally obtained data. If an event is satisfying condition to have BDTG value larger than the cut-off value obtained in the training phase, event will be considered a signal.

Due to a presence of numerous background processes with similar signatures, usually there is no single observable with sufficiently high separating power, but a large number of less sensitive observables should be optimally combined. In Figure 5, we give observables used for classification of $H \rightarrow \gamma\gamma$ events: di-photon energy ($E_{\gamma\gamma}$), di-photon transverse momentum ($p_{T_{\gamma\gamma}}$), di-photon polar angle ($\theta_{\gamma\gamma}$), cosine of the helicity angle ($\cos_{\theta_{hel}}$), transverse momentum of photons (($p_T(\gamma_1)$, $p_T(\gamma_2)$), polar angle of photons (($\theta(\gamma_1), \theta(\gamma_2)$), energy of photons (($E(\gamma_1)$, $E(\gamma_2)$), total E_{ECAL} and E_{HCAL} energy per event - standing for the deposited energies in electromagnetic and hadronic callorimeters, respectively. Di-photon transverse momentum ($p_{T_{\gamma\gamma}}$) turns out to have the largest separation power.

One may observe that a quite large number of observables (12) is used. It is interesting to note that over-enlargement of the number of observables makes MVA ‘biased’ towards the signal sample used in the training phase deteriorating its performance in the application phase. This is known as an over-training. The number of sensitive observables is optimized to keep the values of the Kolmogorov-Smirnov test sufficiently small. Larger values of Kolmogorov-Smirnov test indicates that the method ‘sees’[4] differently the signal in the training and application phase what is a sign of over-training.

[4] The method output observables are compared.

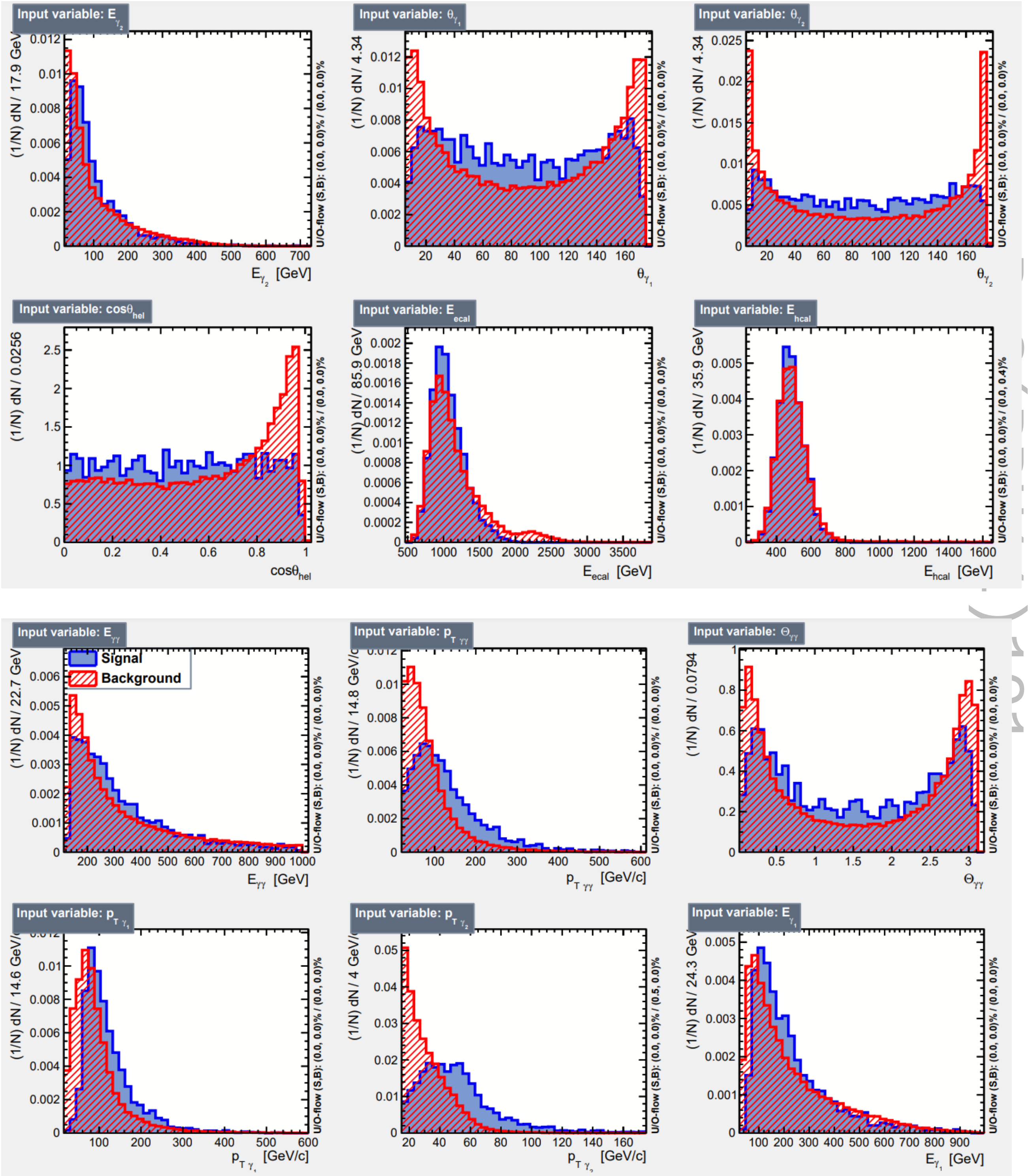

Figure 5. Sensitive observables described in the text used in BDTG training phase for separation of H → γγ signal from background.

4. Discussion

Background suppression before (left) and after (right) MVA application is illustrated in the following processes: $H \rightarrow ZZ^*$ (Figure 6 [5]), $H \rightarrow \gamma\gamma$ (Figure 7 [3]) and $H \rightarrow \mu^+\mu^-$ (Figure 8 [6]). As can be seen, before the application of MVA, background dominates the signal by at least two orders of magnitude. Backgrounds are significantlly suppressed after applying MVA, background to signal ratio of order of $\leq 10^{-1}$ for $H \rightarrow ZZ^*$ (Figure 6, right). The remaining background for rare Higgs decays to $\mu\mu$ and $\gamma\gamma$ is several times larger than the signal despite the fact that it is reduced to a few percent. Signal selection efficiency of MVA are approximatelly: 63%, 53% and 32% for $H \rightarrow \gamma\gamma$, $H \rightarrow ZZ^*$ and $H \rightarrow \mu^+\mu^-$decays respectively. Signal selection efficiency in $H \rightarrow \mu^+\mu^-$ analyses is significantly smaller compared to the $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^*$ analysis. Main reason for this is relatively low number of signal statictics in $H \rightarrow \mu^+\mu^-$ decay at 1.5 ab^{-1}, which is ~ 50 times smaller compared to $H \rightarrow \gamma\gamma$ at 5 ab^{-1} and $H \rightarrow ZZ^*$ at 5 ab^{-1} signal statistics.

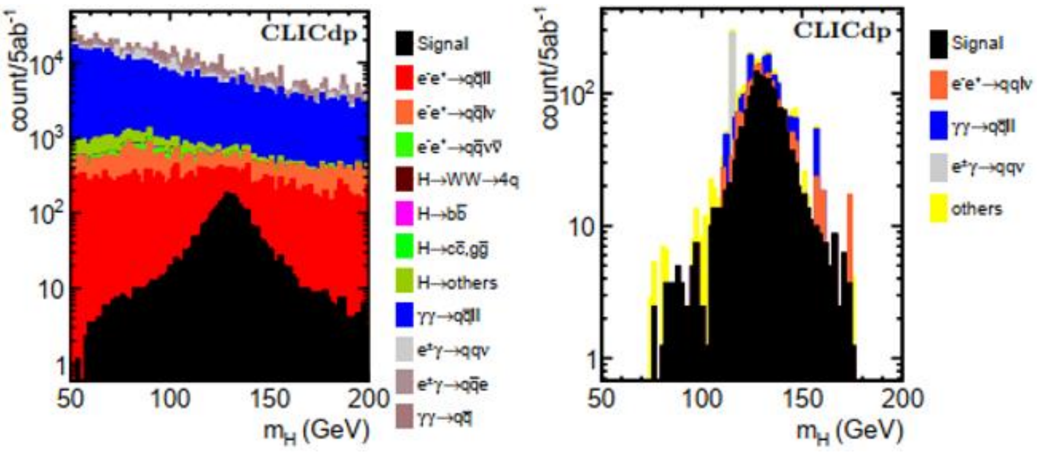

Figure 6. Reconstructed mass of the selected Higgs boson candidate before (left) and after (right) application of MVA selection for signal process $H \rightarrow ZZ^*$ at 3 TeV center-of-mass energy. Background is plotted on top of the signal.

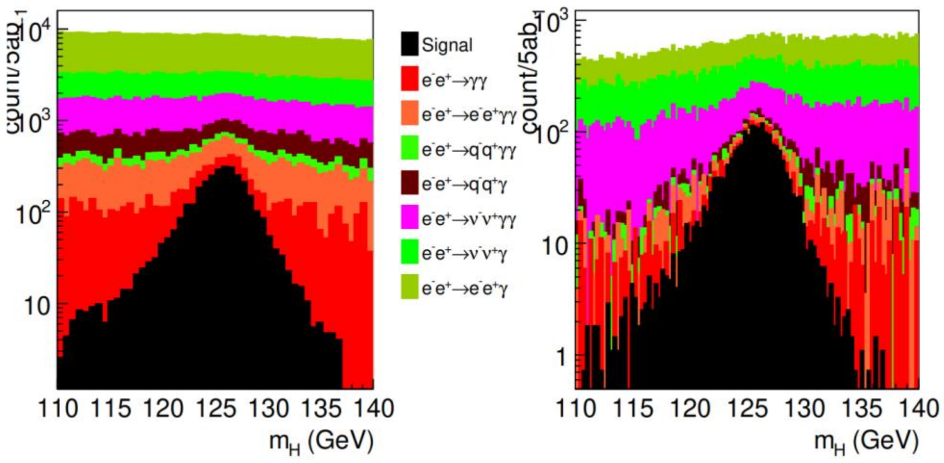

Figure 7. Reconstructed mass of the selected Higgs boson candidate before (left) and after (right) application of MVA selection for signal process $H \rightarrow \gamma\gamma$ at 3 TeV center-of-mass energy. Background is plotted on top of the signal.

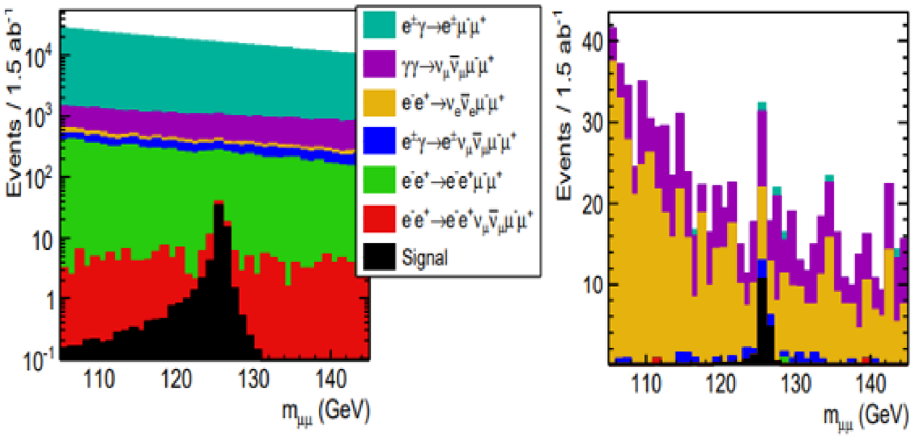

Figure 8. Reconstructed mass of the selected Higgs boson candidate before (left) and after application of MVA selection (right) for signal process $H \to \mu^+\mu^-$ at 1.4 TeV center-of-mass energy. Background is plotted on top of the signal.

In Figures 7 and 8, can be seen that the remaining background dominates over the signal even after applying the MVA, because di-photon invariant mass, which gives the best signal and background separation is not used in these analyses. More details can be found in [3] and [6].

5. Conclusion

The future Higgs factories provide excellent environment to study properties of the Higgs boson. Even the clean environment of future e^+e^- colliders calls for MVA application to separate Higgs signal from the concurrent physics processes. Reaching the precision goals of these experiments (for example, measurement of Higgs couplings at sub-percent level) would not be possible without refined MVA event selection optimized to maximize statistical significance. The need for MVA application in HEP is ever-growing, in particular at future hadron colliders where huge pile-up calls for MVA application already at the event reconstruction level. The possibility to perform precision Higgs physics at the future colliders allows not only to measure Higgs boson properties, but to search for BSM signatures.

6. Acknowledgments

This research was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia and by the Science Fund of the Republic of Serbia through the Grant No. 7699827, IDEJE HIGHTONE-P.

References

[1] N. Vukašinović, I. Bozovic, Higgs Physics at CLIC, 9th International Conference on New Frontiers in Physics, 4-12 September 2020, International Journal of Modern Physics A, 2022 37 (07), 2022 DOI:10.1142/S0217751X22400061.

[2] H. Abramowicz et al., Higgs Physics at the CLIC Electron-Positron Linear Collider. Eur. Phys. J. C 77, 475 (2017). DOI: 10.1140/epjc/s10052-017-4968-5.

[3] Kačarević et al, Measurement of the Higgs boson branching ratio BR (H → γγ) at 3 TeV CLIC, Phys. Rev. D 105, 092009 (2022).

[4] A. Hocker et al., TMVA - Toolkit for multivariate data analysis (2009), arXiv:physics/0703039.

[5] N. Vukašinović et al, Measurement of the H to ZZ branching fraction at a 350 GeV and 3 TeV CLIC, Phys. Rev. D 105, 092008 (2022).

[6] G. Milutinović-Dumbelović et al, Physics potential for the measurement of σ (Hνν) × BR (H → $\mu^+\mu^-$) at 1.4 TeV CLIC collider, CLICdp-Note-2014-005, *Eur. Phys. J. C* **75**, 515 (2015).