Machine learning classification: Case of Higgs boson CP state in $H \rightarrow \tau\tau$ decay at the LHC

K. Lasocha,^{1,2} E. Richter-Was,¹ D. Tracz,¹ Z. Was,³ and P. Winkowska⁴

¹*Institute of Physics, Jagellonian University, Lojasiewicza 11, 30-348 Krakow, Poland*

²*CERN, 1211 Geneva 23, Switzerland*

³*Institute of Nuclear Physics Polish Academy of Sciences, PL-31342 Krakow, Poland*

⁴*Department of Computer Science, AGH USiT, Al. Mickiewicza 30, 30-059 Kraków, Poland*

(Received 22 February 2019; published 3 December 2019)

Machine learning (ML) techniques are rapidly finding a place among the methods of high-energy physics data analysis. Different approaches are explored concerning how much effort should be put into building high-level variables based on physics insight into the problem, and when it is enough to rely on low-level ones, allowing ML methods to find patterns without an explicit physics model. In this paper we continue the discussion of previous publications on the CP state of the Higgs boson measurement of the $H \rightarrow \tau\tau$ decay channel with the consecutive $\tau^\pm \rightarrow \rho^\pm \nu$; $\rho^\pm \rightarrow \pi^\pm \pi^0$ and $\tau^\pm \rightarrow a_1^\pm \nu$; $a_1^\pm \rightarrow \rho^0 \pi^\pm \rightarrow 3\pi^\pm$ cascade decays. The discrimination of the Higgs boson CP state is studied as a binary classification problem between CP even (scalar) and CP odd (pseudoscalar) states using a deep neural network (DNN). Improvements on the classification from the constraints on directly nonmeasurable outgoing neutrinos are discussed. We find that, once added, they enhance the sensitivity sizably, even if only imperfect information is provided. In addition to DNNs we also evaluate and compare other ML methods: boosted trees, random forests, and support vector machines.

DOI: [10.1103/PhysRevD.100.113001](https://doi.org/10.1103/PhysRevD.100.113001)

I. INTRODUCTION

Machine learning (ML) techniques are finding an increasing number of applications in high-energy physics phenomenology. With the Tevatron and the LHC experiments it has become a standard analysis tool. ML techniques are used for event selection, event classification, background suppression for the signal events of the interest, etc. For a recent comprehensive review, see Refs. [1–3]. Over the last years the most significant progress in phenomenology due to ML techniques (in particular the recent developments in neural network methods) has been in hadronic jet reconstruction and classification: jet substructure, flavor, charge, and mass. Some long-standing challenges of more classical algorithms have been addressed; see, e.g., Refs. [4–10].

In this paper we present studies on the seemingly related problem of how the substructure and pattern of hadronically decaying τ leptons can be useful to determine the CP state of the Higgs boson in the decay $H \rightarrow \tau\tau$. The theoretical description of the process including τ -lepton decays is relatively simple and only suffers from minor

theoretical ambiguities. On the other hand, a complete detection approach remains a challenge. For example, indirect constraints had to be devised and validated instead of nonmeasurable τ -neutrino momenta, and the related part of the sensitivity was often compromised.

This problem has a long history [11,12]. It was studied for both electron-positron [13,14] and hadron-hadron colliders [15,16]. Despite some interest, CP states in $H \rightarrow \tau\tau$ decay were not measured or even explored in LHC analysis designs. While more classical experimental analysis strategies have been prepared and documented (see, e.g., Ref. [17]) ML methods for exploring HL-LHC strategies are still at an early stage.

A typical experimental data sample consists of events. Each event can be understood as a point in a multidimensional coordinate space, representing the four-momenta and flavors of observed particles or groups of particles. The physics goal is to identify properties of distributions constructed from these events and to interpret them in a physically meaningful way. The ML algorithms with only low-level features of the event are not necessarily able to efficiently capture all information available. The best-performing strategy still seems to be the mixing of low-level information with human-derived high-level features based on insight into the physics of the problem. Examples of such analyses are presented in Refs. [18,19] where the strategy of mixing low-level and high-level features to

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

remove trivial (physics-wise) symmetries was successfully explored. Then, the ML algorithms do not need to learn some basic physics rules, like rotation symmetry.

In previous papers [20,21] we have demonstrated that ML methods, like deep neural networks (DNNs) [22], can serve as a promising analysis method to constrain the Higgs boson CP state in the decay channel $H \rightarrow \tau\tau$. We considered two decay modes of the τ leptons: $\tau^\pm \rightarrow \rho^\pm \nu$ and $\tau^\pm \rightarrow a_1^\pm \nu$, followed by $\rho^\pm \rightarrow \pi^\pm \pi^0$ and $a_1^\pm \rightarrow \rho^0 \pi^\pm \rightarrow 3\pi^\pm$. This forms three possible hadronic final-state configurations— $\rho^\pm \rho^\mp$, $a_1^\pm \rho^\mp$, and $a_1^\pm a_1^\mp$ —each accompanied by a τ -neutrino pair. The information about the Higgs boson CP state is encoded in the angles between the outgoing decay products and the angles between the intermediate resonance decay planes. In earlier studies [12,23] performed with the rather classical *optimal variable* approach [24], we observed that the best discrimination was achievable from features constructed in the rest frame of the primary intermediate resonance pair of the τ decays, with the z axis aligned with the resonance direction. This idea was also explored in Ref. [20] and will be studied in this paper. We have investigated inputs consisting of mixed low-level and high-level features. Many of the high-level features turned out to be not necessary, but they nevertheless provided benchmark results. On the other hand, (actually simple) nontrivial choices for the representation of some low-level features were necessary to achieve any significant result.

The studies presented in Ref. [20] were limited to input from the hadronic decay products $\pi^\pm$ and π^0 ; no detector effects were taken into account. That study was followed by a more systematic evaluation within the context of experimental analysis [21], namely, applying simplified detector effects to the input features. The conclusions of Ref. [20] on the performance of the DNN method still stand, and we will not repeat this evaluation here.

The studies presented in Ref. [20] showed that the case of $\rho^\pm \rho^\mp$ followed by $a_1^\pm \rho^\mp$ is the most sensitive to the Higgs CP channel, and a somewhat weaker sensitivity is achieved in the $a_1^\pm a_1^\mp$ case. Should all of the decay channels be equally sensitive to the Higgs CP state? In Ref. [26] it was demonstrated that, yes, the sensitivity of each τ decay channel to spin is the same. Unfortunately, this requires control of all momenta of the τ decay products, in particular that of nonmeasurable neutrinos. The studies presented in Ref. [20] did not rely on the complete information, limiting the input information to the hadronic (visible) decay products only. However, it is possible to overcome this limitation and (approximately) reconstruct the neutrino momenta from the τ decay vertex position and event kinematics (the momenta of visible τ decay products, overall missing p_T , and overall collision center-of-mass system energy). Such a reconstruction is challenging from both the experimental and analysis design perspectives: the relations between necessary features are more complicated.

Nevertheless, this provides new opportunities for ML methods, which we will explore with the help of *expert variables*: the azimuthal angles of the neutrino orientation. The possibility that the angle may become experimentally available with adequate precision can be concluded from a recent experimental publication of the LHC collaborations on the measurement of the $H \rightarrow \tau\tau$ signal [27,28], τ substructure reconstruction and classification [29,30], and progress on the precision of B -meson decay vertex position measurements [31–33].

We attempt to reconstruct the two neutrinos' four-momenta (i.e., six quantities) from the experimentally available quantities and examine when such approximate information can be useful. To achieve this goal we propose the following three steps:

- (1) Reconstruct the neutrino 4-momenta components collinear to the directions of the visible decay products of τ leptons, from the missing transverse energy of the event E_{miss}^x , E_{miss}^y and the invariant mass of the Higgs boson m_H .
- (2) Reconstruct the transverse part of the neutrino momenta from the τ -lepton invariant mass constraint.
- (3) Reconstruct the two remaining azimuthal angles ϕ_{ν_1} , ϕ_{ν_2} of the neutrinos (or equivalent information), with the help of τ -decay position vertices.

After step 1 we have four independent variables to constrain, and after step 2 only two independent variables remain. The load on the constraints from the τ decay vertex position (probably the least precisely measured parameter) is minimized. This approach can be understood as an attempt to construct high-level features with the expert-supported design. If useful, this may later be replaced with better choices. Several papers with optimal variables in mind followed such a strategy [13,14,16].

For compatibility, we use the same simulated samples as in Ref. [20], namely, Monte Carlo events of the 125 GeV Higgs boson of the Standard Model, produced in pp collisions at 13 TeV center-of-mass energy, generated with PYTHIA8.2 [34] and with spin correlations simulated using TauSpinner [23]. For τ -lepton decays we use Tauolapp [35]. All spin and parity effects are implemented with the help of the TauSpinner weight wt . This is why the samples prepared for the CP -even or -odd Higgs are correlated. For each channel we use about 10^7 simulated Higgs events [36]. In order to partly emulate detector conditions, a minimal set of cuts is used. We require that the combined transverse momenta of the visible decay products, for each τ , is larger than 20 GeV. We also require that the transverse momentum of each $\pi^\pm$ is larger than 1 GeV.

As in Ref. [20], we perform a DNN analysis for the three channels of the Higgs, i.e., τ -lepton-pair decays, denoted respectively as, $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$. Only two hypotheses on Higgs parity are compared. However, an extension to a parametrized classification (similar to the approach taken in Ref. [37]) could be envisaged as an

obvious next step, e.g., the measurement of the Higgs CP parity-mixing angle. Our paper can also be understood as a work in that direction.

Our baseline for ML methods is the DNN; nonetheless, we also work with more classical ML techniques like boosted trees (BTs) [38], random forests (RFs) [39], and support vector machines (SVMs) [40]. A comparative analysis is presented for the $\rho^\pm - \rho^\mp$ case and for smaller event samples of about 10^6 events.

Our paper is organized as follows. In Sec. II we briefly recall the physics of the problem and the previous results from Ref. [20]. In Sec. III we discuss how to reconstruct (with some approximation) the outgoing neutrino momenta. We exploit the collinear approximation, mass constraints, and information on the spatial positions of the production and decay vertices. In Sec. IV we present an improvement on the DNN classification from information on the neutrinos. We quantify the necessary precision on the neutrinos' azimuthal angle to improve the performance of the classifier. In Sec. V the main results are summarized and an outlook is provided.

In the Appendix A details concerning the implementation of the DNN analysis are given. In Appendix B we present results obtained with the other ML techniques (BTs, RFs, and SVMs). We also discuss technical benchmarks, like CPU usage and transient memory.

II. CLASSIFICATION BASED ON HADRONIC DECAY PRODUCTS

Let us comment briefly on a few selected results [41] from Ref. [20], summarized in Table I. For the DNN classification, only the directly measurable 4-momenta of the hadronic decay products of the τ leptons were considered. They were boosted to the rest frame of the primary intermediate resonance pairs: $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, or $a_1^\pm - a_1^\mp$. All four vectors were later rotated to the frame where the primary resonances are placed along the z axis. This greatly improved the learning process. The DNN algorithm did not have to, e.g., rediscover rotational symmetry, and from the very beginning the internal weights of the DNN algorithms could determine transverse CP -sensitive degrees of freedom from the longitudinal ones. To quantify the performance for Higgs CP classification we used a weighted area under curve (AUC) and receiver

operator characteristic (ROC) curve [42,43]. For each simulated event we know (from the calculated matrix elements) the probability that an event is sampled as a scalar or pseudoscalar (for details, see Appendix A). This forms so-called *oracle predictions*, i.e., the ultimate discrimination for the problem, which is about 0.782, independent [44] of the τ decay channels. Random classification corresponds to 0.500.

For the studied τ -pair decay channels, an AUC in the range 0.557–0.638 was achieved. Note that the AUC score is so much lower than the oracle predictions due to missing information on the neutrino momenta, which are important carriers of spin information but are not directly accessible by measurements. Let us briefly explain the physics context of the problem.

The Higgs boson Yukawa coupling expressed with the help of the scalar-pseudoscalar parity-mixing angle ϕ reads

$$\mathcal{L}_Y = N\bar{\tau}h(\cos\phi + i\sin\phi\gamma_5)\tau, \quad (1)$$

where N denotes normalization, h is the Higgs field, and $\bar{\tau}$ and τ spinors of τ^+ and τ^- . The matrix element squared for the scalar/pseudoscalar/mixed-parity Higgs with decay into $\tau^+\tau^-$ pairs can be expressed as

$$|M|^2 \sim 1 + h_+^i h_-^j R_{i,j}, \quad i, j = \{x, y, z\} \quad (2)$$

where $h_\pm$ denote polarimetric vectors of τ decays (solely defined by the τ decay matrix elements) and $R_{i,j}$ is the density matrix of the τ -lepton pair spin state. Details of the frames used for the definitions of $R_{i,j}$ and $h_\pm$ can be found in Ref. [45]. The corresponding CP -sensitive spin weight wt is simple:

$$wt = 1 - h_+^z h_-^z + h_+^\perp R(2\phi) h_-^\perp. \quad (3)$$

The formula is valid for $h_\pm$ defined in the $\tau^\pm$ rest frames, and h^z and $h^\perp$ stand for the longitudinal and transverse components of h . $R(2\phi)$ denotes the matrix of 2ϕ angle rotation around the z direction: $R_{xx} = R_{yy} = \cos 2\phi$, $R_{xy} = -R_{yx} = \sin 2\phi$. The $\tau^\pm$ decay polarimetric vectors h_+^i , h_-^i , in the simplest case of $\tau^\pm \rightarrow \pi^\pm \pi^0 \nu$ decay, read

$$h_\pm^i = \mathcal{N}(2(q \cdot p_\nu)q^i - q^2 p_\nu^i), \quad (4)$$

TABLE I. The DNN performance taken from Ref. [20] for discrimination between scalar and pseudoscalar Higgs CP states. For DNN classification only the hadronic decay products' 4-momenta were used.

Line content	Channel: $\rho^\pm - \rho^\mp$ $\rho^\pm \rightarrow \pi^\pm \pi^0$	Channel: $a_1^\pm - \rho^\mp$ $a_1^\pm \rightarrow \rho^0 \pi^\pm$, $\rho^0 \rightarrow \pi^+ \pi^-$ $\rho^0 \rightarrow \pi^+ \pi^-$	Channel: $a_1^\pm - a_1^\mp$ $a_1^\pm \rightarrow \rho^0 \pi^\pm$, $\rho^0 \rightarrow \pi^+ \pi^-$
Fraction of $H \rightarrow \tau\tau$	6.5%	4.6%	0.8%
Number of features	24	32	48
Oracle predictions	0.782	0.782	0.782
DNN classification (AUC)	0.638	0.590	0.557

where the τ decay products $\pi^\pm$, π^0 and ν_τ 4-momenta are denoted as $p_{\pi^\pm}$, p_{π^0} , p_ν , and $q = p_{\pi^\pm} - p_{\pi^0}$, respectively. The formula is longer for the $h^\pm_\pm$ of the decay $\tau^\pm \rightarrow \pi^\pm \pi^\pm \pi^\mp \nu$ due to a dependence on the modeling of the decay [21]. Obviously, complete CP sensitivity can be extracted only if p_ν is known. Note that the spin weight wt is a simple first-order trigonometric polynomial in a (doubled) Higgs CP parity-mixing angle. This observation is valid for all τ decay channels.

III. APPROXIMATING COMPONENTS OF NEUTRINO MOMENTA

Our conjecture is that some of the steps listed in the Introduction and presented below may in the future be replaced or optimized by the solutions present in ML libraries. The expert variables—in particular, ϕ_{ν_1} and ϕ_{ν_2} —will not be needed. We first need to explain our construction in detail.

We start with the approximate neutrino momenta in the ultrarelativistic (collinear) approximation. We temporarily assume that the neutrino momenta and visible τ products' momenta are collinear. Later, we relax this simplification. This gives a reasonable approximation for the largest collinear components (in not only the laboratory frame but also the Higgs rest frame and the rest frame of its visible decay products).

A. Collinear approximation

The basic kinematical constraint on the 4-momenta of each $\tau^\pm \rightarrow had^\pm \nu$ decay reads (where $had^\pm$ stands for the combined hadronic system produced in the decay, i.e., $\pi^\pm$, π^0 , etc.)

$$p_{\tau_1} = p_{had_1} + p_{\nu_1}, \quad p_{\tau_2} = p_{had_2} + p_{\nu_2}, \quad (5)$$

where p_{τ_1} , p_{τ_2} denote the 4-momenta of the decaying τ leptons, p_{had_1} , p_{had_2} denote the 4-momenta of their combined hadronic (i.e., measurable) decay products, and p_{ν_1} , p_{ν_2} denote the 4-momenta of the decay neutrinos.

We temporarily assume that the directions of the hadronic decay products and neutrinos are parallel to the direction of the decaying τ and

$$\vec{p}_{had} = x \cdot \vec{p}_\tau, \quad \vec{p}_\nu = (1 - x) \cdot \vec{p}_\tau, \quad (6)$$

where x is in the range (0,1). Then, for τ^+ and τ^- we can write

$$\vec{p}_\nu = \frac{1-x}{x} \cdot \vec{p}_{had} = \alpha \cdot \vec{p}_{had}. \quad (7)$$

From Eq. (7) we obtain

$$|\vec{p}_{\nu_1}| = \alpha_1 \cdot |\vec{p}_{had_1}|, \quad |\vec{p}_{\nu_2}| = \alpha_2 \cdot |\vec{p}_{had_2}|. \quad (8)$$

These relations hold in both the laboratory frame and the rest frame of the hadronic decay products, which is a consequence of the properties of Lorentz transformations of ultrarelativistic particles. This is why we can calculate α_1 and α_2 in the laboratory frame but use them in the rest frame of the combined hadronic decay products. That frame seems to be optimal [20] for the construction of expert variables for ML classification.

1. The E_{miss}^x , E_{miss}^y constraints

The laboratory-frame event momentum imbalance in the plane transverse to the beam direction, usually denoted as E_{miss}^x , E_{miss}^y , can be used to constrain neutrino momenta. It can be attributed to the sum of the transverse components of the neutrino momenta, but it also accumulates all imperfections of the reconstruction of the other outgoing particles of that event. Then, thanks to Eq. (7),

$$\begin{aligned} E_{\text{miss}}^x &= p_{\nu_1}^x + p_{\nu_2}^x = \alpha_1 \cdot p_{had_1}^x + \alpha_2 \cdot p_{had_2}^x, \\ E_{\text{miss}}^y &= p_{\nu_1}^y + p_{\nu_2}^y = \alpha_1 \cdot p_{had_1}^y + \alpha_2 \cdot p_{had_2}^y, \end{aligned} \quad (9)$$

and

$$\alpha_1 = \frac{E_{\text{miss}}^x - \alpha_2 \cdot p_{had_2}^x}{p_{had_1}^x} \quad (10)$$

or

$$\alpha_1 = \frac{E_{\text{miss}}^y - \alpha_2 \cdot p_{had_2}^y}{p_{had_1}^y}. \quad (11)$$

Finally, by solving for α_1 and α_2 we obtain the expressions

$$\begin{aligned} \alpha_2 &= \frac{E_{\text{miss}}^y \cdot p_{had_1}^x - E_{\text{miss}}^x \cdot p_{had_1}^y}{p_{had_2}^y \cdot p_{had_1}^x - p_{had_2}^x \cdot p_{had_1}^y}, \\ \alpha_1 &= \frac{E_{\text{miss}}^x - \alpha_2 \cdot p_{had_2}^x}{p_{had_1}^x}, \end{aligned} \quad (12)$$

which are useful for studies of ML classification.

2. Using the m_H constraint

Equation (12) alone provides solutions for α_1 and α_2 . However, E_{miss}^x , E_{miss}^y have large experimental uncertainties. At the same time, the high-quality constraint from the known Higgs-boson and τ -lepton masses is

$$\begin{aligned} m_H^2 &= (p_{\tau_1} + p_{\tau_2})^2 \\ &= 2 \cdot m_\tau^2 + 2 \cdot (1 + \alpha_1) \\ &\quad \cdot (1 + \alpha_2) [E_{had_1} E_{had_2} - p_{had_1}^x \cdot p_{had_2}^x \\ &\quad - p_{had_1}^y \cdot p_{had_2}^y - p_{had_1}^z \cdot p_{had_2}^z]. \end{aligned} \quad (13)$$

E_{had_1} and E_{had_2} denote the energies of the hadronic systems had_1 and had_2 . Later, we will use the similar notation E_ν for the neutrino energy.

Unfortunately, only the product $(1 + \alpha_1) \cdot (1 + \alpha_2)$ can be controlled in this way,

$$(1 + \alpha_1) \cdot (1 + \alpha_2) = \frac{m_H^2/2 - m_\tau^2}{E_{had_1} E_{had_2} - p_{had_1}^x \cdot p_{had_2}^x - p_{had_1}^y \cdot p_{had_2}^y - p_{had_1}^z \cdot p_{had_2}^z}. \quad (14)$$

3. Choosing an optimal solution for the longitudinal neutrino momentum

To constrain α_1 and α_2 , we have the three independent equations [Eqs. (12) and (14)] at our disposal. We have checked that all three options lead to comparable predictions and marginal differences in the ML performance, at least as long as the measurement ambiguities of E_{miss}^x , E_{miss}^y are not taken into account:

- (i) *Approx-1*: Eq. (12) only.
- (ii) *Approx-2*: Eq. (14) and α_1 from Eq. (12).
- (iii) *Approx-3*: Eq. (14) and α_2 from Eq. (12).

The actual choice may optimize experimental precision. For now, the option *Approx-1* is chosen as a baseline for the results [46] without much elaboration.

To illustrate the effectiveness, the correlation between α_1 -true [47] and α_1 -*Approx-1* is shown for the $a_1^\pm - \rho^\mp$ case in the top panel of Fig. 1. In the bottom panel, as a consistency check, the correlation of the $a_1^\pm - \rho^\mp$ rest-frame and laboratory-frame energy fraction x_1 calculated using α_1 -*Approx-1* is given. A sample of 10^4 events was used for these scatter plots. The fraction of events contained in the band $\Delta\alpha_1/\alpha_1 = \pm 5\%(\pm 10\%)$ is about 25%(39%), and the fraction in the band $\Delta x_1/x_1 = \pm 1\%$ is about 85%. This relatively poor resolution in α_1 will be reflected in the resolution of the approximate neutrino momenta. It will be interesting to observe how much it will affect the classification capability of trained DNNs, which will be discussed in Sec. IV.

B. Energy and transverse components of neutrino momenta

Now, with the help of the approximate p_ν^z (aligned with the direction of combined visible decay products), we can turn our attention to p_ν^x and p_ν^y . In the rest frame of the hadronic decay product system the $p_{had_{1,2}}$ momenta are set along the z direction, and thus $p_{had}^x = p_{had}^y = 0$. The τ mass constraint reads

$$m_\tau^2 = (E_\nu + E_{had})^2 - (p_\nu^x)^2 - (p_\nu^y)^2 - (p_\nu^z + p_{had}^z)^2, \quad (15)$$

and for massless ν_τ

FIG. 1. Top: Correlation between the true and approximate α_1 calculated for $a_1^\pm - \rho^\mp$. Bottom: Correlation between the fraction x of the τ -lepton momentum carried by the hadronic decay products obtained in the *Approx-1* approximation, in the $a_1^\pm - \rho^\mp$ and laboratory frames.

$$0 = (E_\nu)^2 - |p_\nu^T|^2 - (p_\nu^z)^2. \quad (16)$$

The equations lead to the following relations:

$$E_\nu = \frac{m_\tau^2 - E_{had}^2 + (p_{had}^z)^2 + 2 \cdot p_\nu^z p_{had}^z}{2E_{had}}, \quad (17)$$

$$p_\nu^T = \sqrt{E_\nu^2 - (p_\nu^z)^2},$$

where for $p_\nu^z = \alpha \cdot p_{had}^z$ one of the α approximations from Sec. III A 3 is used.

α_1 , α_2 , E_{ν_1} , and E_{ν_2} must be positive; otherwise, the approximation fails and the event cannot be used. Also, events with a negative approximate $(p_\nu^T)^2$ could be rejected, but for our studies we instead decide to set this component to zero. In total, about 17% of events are rejected for *Approx-1*. Further 11% of events are rejected if the above criteria are fulfilled with *Approx-2* and *Approx-3* too.

FIG. 2. The $(x_{\text{true}} - x_{\text{approx.}})/x_{\text{true}}$ distribution for the reconstructed neutrino energy (top), and longitudinal (middle) and transverse (bottom) momenta using *Approx-1*. Events of shifts outside the window $(-1.0, 1.0)$ are collected in the -1.0 bin. The standard deviation (denoted as RMS) is calculated for the range $(-1.0, 1.0)$.

In Fig. 2 we show the distribution of the relative shifts from the generated to approximate E_{ν} , p_{ν}^z and p_{ν}^T for the $a_1^{\pm} - \rho^{\mp}$ case. The p_{ν}^T is approximated better than E_{ν} and p_{ν}^z .

We remain encouraged because for ML classifications even approximate observables (expert variables) may be useful to improve classification scores.

C. Azimuthal angles of neutrinos

After completing step B, we are left without two azimuthal angles for the orientation of $p_{\nu_1}^T$ and $p_{\nu_2}^T$ only. To capture the sensitivity of the Higgs boson CP they have to be known, preferably in the visible τ -pair decay products rest frame. Those two angles can be inferred from the positions of the τ decay vertices and then through boosts and rotations related to the azimuthal angles in the visible decay product frame.

The transverse coordinates of the primary interaction point are to a good precision consistent with zero. At the same time, the tracks of the τ decay products will not point to this interaction vertex, but rather to the position of the τ decay vertex shifted by the τ 's flight path. The direction of the τ 's flight path and (as a consequence) its momentum components can be reconstructed. This provides a constraint on the ν_{τ} momentum as well. We do not intend to go into the details of this challenging secondary vertex position measurement. We refer to Refs. [31,32], which discuss the similar problem of the secondary vertex in the case of B -meson decay and its application to the classification of hadronic jets. One may assume that such a measurement is possible for a τ lepton, and that the orientation of the ν_{τ} momentum around the direction of the visible hadronic τ decay products can be constrained.

To determine how precisely we need to know this information, we take the true azimuthal angles ϕ_{ν_1} and ϕ_{ν_2} in the rest frame of the visible decay products and smear them. For the $\Delta\phi_{\nu} = |\phi_{\nu}^{\text{smear}} - \phi_{\nu}^{\text{true}}|$ smearing probability we take

$$f_{\text{train}}(\Delta\phi_{\nu}, \beta) = \frac{1}{\beta} \exp\left(-\frac{1}{\beta} \Delta\phi_{\nu}\right). \quad (18)$$

We have chosen the exponential shape instead of the often-used Gaussian shape [48]. Note however that the length of the τ flight path follows an exponential distribution. We choose the sign for the shift with equal probabilities.

We think that at present it is premature to attempt a realistic detector smearing. Such attempts to investigate experimental smearings for the secondary vertex position have only been reported for the case of the $\tau^{\pm} \rightarrow \pi^{\pm} \pi^{\pm} \pi^{\mp} \nu$ decay channel of $Z/\gamma^* \rightarrow \tau\tau$ production at the LHC [49].

D. Ansatz for the direction of the τ leptons

In Sec. III C we discussed the possibility of adding approximate information on the angle of the outgoing neutrino in the decay plane to the feature list. However, for the multivariate methods this angle does not have to be explicitly present in the feature list. In fact, indirect

TABLE II. Lists of features for ML classification, marked as Variant-X.Y. In the third column, the number of features for the $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ channels are given. All components of the 4-momenta are taken in the hadronic decay product rest frame. The primary resonances ($\rho^\pm$, $a_1^\pm$) are aligned with the z axis. E_{miss}^x and E_{miss}^y are in the laboratory frame. In practice, instead of p_ν^T and ϕ_ν , the pair of variables $p_\nu^T \cos \phi_\nu$ and $p_\nu^T \sin \phi_\nu$ is used.

Notation	Features	Counts	Comments
Variant-All	4-momenta ($\pi^\pm$, π^0 , ν)	24/28/32	
Variant-1.0	4-momenta ($\pi^\pm$, π^0)	16/20/24	as in Table 3 of Ref. [20]
Variant-1.1	4-momenta ($\pi^\pm$, π^0 , $\rho^\pm$, $a_1^\pm$), m_i^2 , m_k^2 , y_i , y_k , $\phi_{i,k}^*$	29/46/94	
Variant-2.0	4-momenta ($\pi^\pm$, π^0), E_ν , p_ν^z , p_ν^T	22/26/30	
Variant-2.1	4-momenta ($\pi^\pm$, π^0), E_ν , p_ν^z , p_ν^T	22/26/30	Approx. E_ν , p_ν^z , p_ν^T
Variant-2.2	4-momenta ($\pi^\pm$, π^0), E_ν , p_ν^z , p_ν^T , E_{miss}^x , E_{miss}^y	24/28/32	Approx. E_ν^z , p_ν^z , p_ν^T
Variant-3.0.0	4-momenta ($\pi^\pm$, π^0), E_ν , p_ν^z , p_ν^T , ϕ_ν	24/28/32	Approx. E_ν , $\vec{p}_\nu$
Variant-3.1. β	4-momenta ($\pi^\pm$, π^0), E_ν , p_ν^z , p_ν^T , ϕ_ν	24/28/32	Approx. E_ν , $\vec{p}_\nu$; ϕ_ν smeared with β
Variant-4.0	4-momenta ($\pi^\pm$, π^0 , $\tau^\pm$)	24/28/32	
Variant-4.1	4-momenta ($\pi^\pm$, π^0 , $\tau^\pm$)	24/28/32	Approx. p_τ

information such as the approximate direction of the outgoing τ lepton may be good enough.

From the primary and secondary vertex positions, the direction of the laboratory system τ -lepton momentum, i.e., p_x , p_y , p_z is constrained. Assuming a known τ time of flight (t_{flight}) and mass m_τ , we calculate

$$p_i^\tau = m_\tau \cdot (i_{\text{sec vtx}} - i_{\text{prim vtx}}) / t_{\text{flight}}, \quad (19)$$

where $i_{\text{sec vtx}}$ and $i_{\text{prim vtx}}$ denote the spatial position of the reconstructed primary and secondary vertex, respectively, in the laboratory (collision) frame ($i = x, y, z$). Instead of the unknown true time of flight, we use the value from the Particle Data Group, $c\tau_\tau = 87 \mu\text{m}$ [50]. The true time of flight behaves according to the exponential distribution with mean $\langle t_{\text{flight}} \rangle = \tau_\tau$. This imposes that the approximation used to estimate p_x , p_y , and p_z is also characterized by an exponential distribution, with mean and sigma close to their true values. The energy of the τ lepton is then calculated using the τ mass constraint,

$$E^\tau = \sqrt{(p_x^\tau)^2 + (p_y^\tau)^2 + (p_z^\tau)^2 + m_\tau^2}. \quad (20)$$

Now, the complete 4-momentum of each τ is boosted into the $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, or $a_1^\pm - a_1^\mp$ system rest frame and added to the feature lists for DNN training.

IV. CLASSIFICATION WITH DNNs

The structure of the data and neural network architecture follows Ref. [20]. We start from the code used there. For the convenience of the reader, we summarize the technical description of our DNN model in Appendix A.

Simulated data consist of events where all decay products are stored together with their flavors. The 4-momenta of the laboratory frame are stored and (whenever they are

needed) transformed to their respective rest frames, as explained in Sec. II. With respect to the analysis published in Ref. [20] we explore approximate information on neutrino momenta derived from the kinematical constraints of the Higgs decay products. We show that significant improvement may originate from even very inaccurate information on the azimuthal angles of the neutrinos' directions.

We explore the potential of classification with the DNN technique with several variants of the feature lists, as detailed in Table II. They are grouped and marked as Variant-X.Y, where X labels the choice of the main features and Y (in most cases) labels if they are calculated from the generator-level 4-momenta or from the approximation; it may also mark if additional, high-level variables are used. It gives us very useful tools to quantify how much of the DNN performance we are losing due to certain approximations made on the groups of features.

In Table III, we collect the AUC scores and average precision (APS) scores [51] obtained on the test sample of simulated data (i.e., events not used for training or validation) with the DNN trained on 50 epochs and with a dropout = 0.20. Both are comparable, with the APS scores being systematically slightly lower, except for a few cases of the $a_1^\pm - a_1^\mp$ channel. This configuration was found to be the most stable for the comparison of Variant-X.Y classifications, but it does not necessarily represent the optimal performance of the particular variant of the feature list. In the first line of Table III we recall the oracle predictions [52]; for details, see Appendix A. It cannot be outperformed by the DNN of any Variant-X.Y. It may not be reached even with a feature list containing the complete set of 4-momenta of τ decay products, denoted as Variant-All.

In the following subsections we discuss these results in detail.

TABLE III. The AUC and APS scores to discriminate scalar and pseudoscalar CP states of the Higgs boson, obtained on the test sample. The DNN was trained on 50 epochs with a dropout of 0.2 (except for the explicitly marked case of Variant-All). Results for the $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ channels are given. The first column labels the choice of features. For details, see Table II.

Features list	AUC/APS ($\rho^\pm - \rho^\mp$)	AUC/APS ($a_1^\pm - \rho^\mp$)	AUC/APS ($a_1^\pm - a_1^\mp$)
Oracle predictions	0.784/0.785	0.781/0.783	0.780/0.782
Variant-All (drop = 0.0)	0.784/0.786	0.778/0.778	0.773/0.774
Variant-All	0.769/0.764	0.748/0.742	0.728/0.720
Variant-1.0	0.655/0.654	0.603/0.602	0.573/0.578
Variant-1.1	0.656/0.655	0.609/0.607	0.580/0.585
Variant-2.0	0.663/0.663	0.626/0.625	0.594/0.595
Variant-2.1	0.664/0.666	0.622/0.622	0.591/0.593
Variant-2.2	0.664/0.666	0.622/0.622	0.591/0.593
Variant-3.0.0	0.771/0.771	0.749/0.743	0.728/0.721
Variant-3.1.2	0.760/0.759	0.738/0.730	0.718/0.710
Variant-3.1.4	0.738/0.735	0.714/0.705	0.687/0.677
Variant-3.1.6	0.715/0.713	0.689/0.680	0.660/0.652
Variant-4.0	0.769/0.766	0.748/0.742	0.728/0.720
Variant-4.1	0.738/0.733	0.704/0.696	0.683/0.676

A. Benchmarks using all or only hadronic decay products

For the first benchmark each event is represented by the 4-momenta of both τ -leptons' decay products (including neutrinos) in the rest frame of all hadronic decay products. This set of features is denoted as Variant-All. The results are displayed in the second and third lines of Table III. The DNN should be able to reproduce oracle predictions, which is almost the case if the dropout is not used, but it only approaches them for the baseline configuration with a dropout of 0.20. The dropout lowers the DNN's performance with Variant-All, but we have verified that for other feature lists this is not always the case. It helps to suppress overfitting, as illustrated in Fig. 8 of Appendix A. In the top panel of Fig. 3 we show for the $a_1^\pm - \rho^\mp$ channel Variant-All, the AUC score as a function of the number of epochs used for training and validation for the $a_1^\pm - \rho^\mp$ channel Variant-All. Scores up to about 0.75 are reached for the validation sample and Variant-All.

For the second benchmark, following Ref. [20], the same events are used but with the features limited to the 4-momenta of visible τ leptons decay products and quantities derived directly from them [53]. The set with only the 4-momenta of visible decay products in the respective rest frames of intermediate resonances is called Variant-1.0. When this is supplemented with higher-level expert features, like the invariant masses of intermediate resonances or energy fractions, it is called Variant-1.1. For all three channels, the results for Variant-1.0 and Variant-1.1 are similar.

Expert variables provide redundant information only. In Fig. 3 (top plot) AUC results for training and validation of $a_1^\pm - \rho^\mp$ are shown for Variant-1.0. The highest result on the validation sample is around 0.60.

In the bottom panel of Fig. 3 we show the ROC curves in the true positive rate (TPR) versus false positive rate (FPR) plane for Variant-All and Variant-1.0.

The achieved AUC and APS scores are collected in Table III. The large difference in the AUC and APS performance between the Variant-All and Variant-1.0 feature sets is present for all channels. In the following we attempt to improve the performance using information on the neutrino momenta and in particular their azimuthal angles.

B. Adding neutrino momenta

In this subsection we present improvements due to the energy and longitudinal neutrino momenta. Such an extension of the feature list is not expected to be very beneficial, as CP information is carried by the transverse degrees of freedom, but it may optimize the use of information learned from the correlations of hadronic decay products.

With the assumptions explained in Sec. III, we approximate each of the neutrino momentum components E_ν , p_ν^z , and p_ν^T in the rest frame of hadronic decay products. It is interesting to first check what is the potential impact of that information, i.e., when truth-level values are used. We add the laboratory-frame E_{miss}^x , E_{miss}^y , redundant to some extent, as it was already used in Eq. (7) for p_ν^z .

The augmented feature list, using the true components of neutrino momenta, is denoted as Variant-2.0, while

FIG. 3. The AUC score (top panel) for training and validation of $a_1^\pm - \rho^\mp$ samples, as a function of the number of epochs and ROC curve for 50 epochs (bottom panel). The feature lists Variant-All and Variant-1.0 were chosen. The training and validation curves overlap.

the ones using the approximate components of neutrino momenta are denoted as Variant-2.1 and Variant-2.2, depending on whether the information on E_{miss}^x and E_{miss}^y is included. The AUC and APS scores from the DNN for the $\rho^\pm - \rho^\mp$, $a_1^\mp - \rho^\pm$, and $a_1^\pm - a_1^\mp$ channels are displayed in Table III. The improvement from Variant-1.0 to Variant-2.0 is not impressive. We also observe a small performance loss from Variant-2.0 to Variant-2.1, which uses approximate neutrino features, resulting in sensitivity loss. The laboratory-frame E_{miss}^x and E_{miss}^y of Variant-2.2 are, as expected, of no help. In Fig. 4 we show the DNN performance for the $a_1^\pm - \rho^\mp$ samples as a function of the number of training epochs: the AUC score achieved as a function of the number of epochs and the ROC curves.

All three approximations for E_ν , p_ν^z , and p_ν^T are studied for the feature sets Variant-2.1 and Variant-2.2.

FIG. 4. Top: AUC score for the training and validation of $a_1^\pm - \rho^\mp$ samples, as a function of the number of epochs. The feature lists Variant-2.0 and Variant-2.1 were chosen. Bottom: Corresponding ROC curves for 50 epochs.

The differences between *Approx-1*, *Approx-2*, and *Approx-3* are small but will certainly be apparent once detector effects are included.

Clearly, the improvement from the approximate information on the neutrino energy and momenta (longitudinal component and overall size of transverse components) is rather small for all three channels. The most sensitive information on the CP state lies in the azimuthal angles of the individual neutrinos, that is, in the individual p_ν^x, p_ν^y components of the hadronic decay products rest frame and not in $p_\nu^T = \sqrt{(p_\nu^x)^2 + (p_\nu^y)^2}$. Realistically, any information on the individual p_ν^x, p_ν^y could be reconstructed only if the measurement of the τ decay vertices was possible. In the next section, we evaluate how accurately this information has to be known to become useful. This constitutes a separate experimental challenge. Note that at this step all components of the ν_τ momenta except the individual p_ν^x, p_ν^y are reconstructed sufficiently well from the measurable quantities.

C. Azimuthal angles of neutrinos from decay vertices

The azimuthal angles ϕ_{ν_1} and ϕ_{ν_2} can be obtained from the measurement of the τ -lepton decay vertices. This allows to reconstruct the τ -lepton momenta and hopefully can be used for our purpose as well. This is a rather widely used technique in the experimental measurements (see, e.g., Ref. [54]), but so far it has only been used for τ -mass and τ -lifetime measurements rather than for neutrino azimuthal angles.

We do not aim to reconstruct these angles; instead, we simply calculate them from the neutrino 4-momenta and add Variant-3.0 and Variant-3.1. β to the feature lists [55]. The first includes the true $\phi_{\nu_1}^{\text{true}}$ and $\phi_{\nu_2}^{\text{true}}$ are used, and the second includes the smeared $\phi_{\nu_1}^{\text{smeared}}$ and $\phi_{\nu_2}^{\text{smeared}}$. In Fig. 5 the $\phi_{\nu}^{\text{true}} - \phi_{\nu}^{\text{smeared}}$ distribution for $\beta = 0.4$ of Eq. (18) is shown.

The AUC scores are evaluated for β in range (0,2). In Fig. 6 the AUC scores for test samples of the three channels are given as a function of β . The AUC scores for $\beta = 0.0$ reproduce (as they should) those of Variant-3.0 and are not very far from the scores of Variant-All. This is because the only difference is the approximate information on the energy and longitudinal and transverse momenta of the neutrino. For β above 1.4, the AUC scores decrease to those of the Variant-2.1 sets, which are then equivalent to not having information on the neutrino azimuthal angles at all. Even $\phi_{\nu_1}^{\text{smeared}}$ and $\phi_{\nu_2}^{\text{smeared}}$ corresponding to the rather large $\beta = 0.4$ contribute sizably to CP Higgs sensitivity. The derivative of the sensitivity with respect to β reaches its maximum at about 0.35 and remains constant until $\beta = 0.9$. Then, nearly all sensitivity gain is lost. For even larger β , the loss of sensitivity continues, but as the contribution is then already small, the deterioration is small too.

FIG. 5. The (true-smeared) distribution of $\Delta\phi_{\nu_1}$ for $\beta = 0.4$ with/without the additional polynomial modulation of Eq. (21)/Eq. (18).

FIG. 6. The AUC score as a function of the smearing parameter β of Eq. (21) for the $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ channels and Variant-3.1. β .

Let us now check if the DNN algorithm is sensitive to the precise modeling of the $\phi_{\nu_{1,2}}$ resolution. This is why, for the validation and test sample, we introduce [56] an additional polynomial component for the smearing,

$$f_{\text{valid}}(\Delta\phi_\nu, \beta, b, c) = f_{\text{train}}(\Delta\phi_\nu, \beta)(1 + b^2\Delta\phi_\nu^2 + c^2\Delta\phi_\nu^4). \quad (21)$$

The results should mimic the impact of inefficiencies (mismodeling) of the DNN training sample with respect to what is present in the validation or test samples. In Fig. 5 the distribution of $\phi_{\nu}^{\text{true}} - \phi_{\nu}^{\text{smeared}}$ is given for $\beta = 0.4$, and $b, c = 0.3, 0.8$. In Table IV we collect the results for the

TABLE IV. The AUC and APS scores for the $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ channels with the feature list Variant-3.1. β ($\beta = 0.2, 0.4$, and 0.6 are used for training, validation, and test samples). For the test sample, polynomial modifications of the smearing function (21) were introduced.

Parameters	AUC/APS		
	($\beta = 0.2$)	($\beta = 0.4$)	($\beta = 0.6$)
$\rho^\pm - \rho^\mp$			
$b = 0.0, c = 0.0$	0.761/0.759	0.739/0.737	0.715/0.714
$b = 0.3, c = 0.8$	0.760/0.758	0.739/0.736	0.716/0.713
$b = 0.9, c = 0.9$	0.759/0.756	0.738/0.734	0.714/0.713
$a_1^\pm - \rho^\mp$			
$b = 0.0, c = 0.0$	0.739/0.731	0.714/0.706	0.687/0.679
$b = 0.3, c = 0.8$	0.738/0.730	0.714/0.705	0.687/0.679
$b = 0.9, c = 0.9$	0.737/0.728	0.714/0.704	0.687/0.678
$a_1^\pm - a_1^\mp$			
$b = 0.0, c = 0.0$	0.713/0.705	0.690/0.680	0.660/0.653
$b = 0.3, c = 0.8$	0.715/0.706	0.693/0.682	0.661/0.653
$b = 0.9, c = 0.9$	0.714/0.706	0.688/0.680	0.660/0.653

$\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ channels and for $\beta = 0.2, 0.4, 0.6$ with further choices of b and c . The additional polynomial component of the smearing introduced to the test sample does not affect the DNN performance. We can see that the degradation due to $b, c = 0.3, 0.8$ is small and the results provide some encouraging insight into the DNN's capacity to exploit imprecise information and point to a possible direction for studies of systematic uncertainties [57].

In our study, when the precision of the experimental inputs was expected to be better than that from the decay vertex impact parameters, we reconstructed the neutrino momenta components from hadronic products and conservation laws. Only the ϕ_ν angles required this rather low-precision input. From Fig. 6 we can expect that the approximate ϕ_ν angle with an ambiguity of up to $\pm \frac{\pi}{4}$ may sizably improve sensitivity.

Such a conjecture on the size of ϕ_ν smearing that is critical for CP sensitivity is of interest for any ML application. For $\beta = 1.2$ the shift $\Delta\phi_\nu$ is bigger than $\frac{\pi}{4}$ in a sizable fraction of events. In this case, the DNN solution does not gain sensitivity from ϕ_ν . Still, an approach relying less on the ϕ_ν measurement and more on restricting which events should be dropped from the analysis could be useful. For large smearing, the elimination of events with a high risk of ϕ_ν misreconstruction may be appropriate, as was attempted in Ref. [14]. Discussion of physics properties choice of the ML algorithms may be then of interest.

D. Tau lepton direction

The approximate information on the τ -lepton direction enables the DNN to constrain the neutrinos and significantly improve the classification. For this purpose, Variant-4.0 and Variant-4.1 are defined in Table II.

In Table III, the performances of the DNN are presented, when the true-level or approximate τ -lepton spatial momenta components in the $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ rest frames are added. We observe a significant improvement of the performance with respect to Variant-2.1 and a comparable performance to the Variant-3.1.X family. In fact, the performance of Variant-4.0 is close to that of Variant-All. Variant-4.1 is a bit lower, close to Variant-3.1.4. Then, only the τ direction in the laboratory frame is exact and the energy is obtained from the simple ansatz of Sex. III D. When such τ 4-momentum is boosted to the $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ rest frames, its direction absorbs some biases. The results of Variant-4.1 indicate that the DNN efficiently converts such an input into information on ν_τ .

V. SUMMARY

From the perspective of theoretical modeling, the CP -parity phenomenology in the cascade decay $H \rightarrow \tau\tau$,

$\tau^\pm \rightarrow had^\pm \nu_\tau$ is rather simple, because the matrix elements can be easily defined. On the other hand, the parity effect manifests itself in the rather complicated features of multidimensional distributions, where kinematic constraints related to ultrarelativistic boosts and detection ambiguities play an important role in the reconstruction of the τ decay kinematics. Our aim was to evaluate precision requirements for experimental features to become useful.

In our previous paper [20] we studied the performance of the DNN binary classification technique for the hadronic τ -lepton decay products only. In this paper we turned our attention to the ν_τ momenta.

Whenever possible, we exploited constraints on the τ mass, H mass, and energy-momentum conservation to minimize dependence on the highly smeared neutrino kinematics deduced from the impact parameters of τ decay and production vertices. The resulting set of expert variables helps DNN algorithms to identify physics-sensitive variables that are useful for identifying differences between the event classes.

Reconstructed with the help of approximation (but only from the visible decay products) longitudinal components of the neutrino momenta alone improved the AUC scores from 0.656, 0.609, and 0.580 to about 0.664, 0.622, and 0.591 for the $\rho^\pm - \rho^\pm$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ cases respectively. The improvement for the Higgs boson CP sensitivity is rather minuscule, even when the detector effects are not taken into account.

A more significant improvement came when the transverse components of the neutrino momenta were known, even imprecisely. This can be achieved if the τ -lepton decay vertices are measured and used to reconstruct the directions of the τ -lepton momenta. The performance of such a reconstruction is detector specific and is a challenge. We have estimated how big of an improvement in CP sensitivity is obtained as a function of detection smearing for the azimuthal angles ϕ_ν and $\phi_{\bar{\nu}}$. Even with a large smearing of $\beta = 0.4$, the AUC scores improved from 0.664, 0.622, and 0.591 to about 0.738, 0.714, and 0.687 for the $\rho^\pm - \rho^\pm$, $a_1^\pm - \rho^\mp$ and $a_1^\pm - a_1^\mp$ cases, respectively. Note that ϕ_ν and $\phi_{\bar{\nu}}$ represent an intermediate step in the transition from expert variables to DNN algorithms with the direct use of low-level features. We leave the topic of the angle measurements and their use to future work.

Similar performance is expected when high-quality τ -lepton laboratory-frame direction data (as seen in the rest frame of all visible Higgs decay products) is available for the evaluation of the τ direction. The ambiguity in the laboratory-frame τ energy is not that important. An enhancement in the $\tau^\pm$ directions was achieved (Variant-4.1), and the AUC scores reached 0.738, 0.704, and 0.683 for the $\rho^\pm - \rho^\pm$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ cases, respectively.

FIG. 7. The ROC curves for different feature lists.

In Fig. 7 we show the ROC curves for the different feature lists discussed in this paper.

The concept of *optimal observables* has been used for many years to obtain phenomenologically sound results. It provides essential tests for ML classification, where multi-dimensional input is used. An approach where sophisticated methods are used to measure $h_{\pm}$ of Eq. (2) should be mentioned. All of the complexity of the hadronic τ decays and detector response is then hidden in each $\tau^{\pm}$ polarimetric vector $h_{\pm}$. Once an algorithm for $h_{\pm}$ reconstruction is obtained, the latter step of CP phenomenology is straightforward: the details of the $\tau^{\pm}$ decay channels and detector effects are resolved. The $h_{\pm}$ complexity is smaller than that of the entire $H \rightarrow \tau\tau$ cascade decay. It is independent of the Higgs phenomenology and calculations can rely on the much more abundant $Z \rightarrow \tau\tau$ data. Such a possibility was mentioned in Ref. [58] and is being pursued by, e.g., the CMS Collaboration. Then, ML techniques could be used to reconstruct $h_{\pm}$ vectors from the complex detector responses to particular τ decay channels and the details of their decay vertex positions.

The evaluation of which of the methods is best, or even how complementary the methods can be, requires the work of experimental groups.

Recently, in Ref. [59] classifiers specifically tuned to tackle the Lorentz group features of high-energy physics signatures were prepared and used. This could be useful for Variant-1.0, where only the 4-momenta of secondary $H \rightarrow \tau\tau$ decay products were used. In the present work this may be less straightforward some of the features are intimately related to the laboratory frame and their transformation to other frames may be poorly defined. This is why expert-variable-style reconstruction of neutrino azimuthal angles may be an efficient course to follow, or at least useful to better understand the limitations and ambiguities of these methods.

ACKNOWLEDGMENTS

P. W. would like to thank L. Grzanka for valuable comments and suggestions during the preparation. This project was supported in part by funds from the Polish National Science Centre under decisions DEC-2017/27/B/ST2/01391. D. T. and P. W. were supported by funds from the Polish National Science Centre under decisions DEC-2014/15/B/ST2/00049. The majority of the numerical calculations were performed at the PLGrid Infrastructure of the Academic Computer Centre CYFRONET AGH in Krakow, Poland.

APPENDIX A: DEEP NEURAL NETWORK

The structure of the simulated data and the DNN architecture follows that published in our previous paper [20]. It is prepared for TensorFlow [60], an open-source machine learning library. The learning procedure is optimized using a variant of the stochastic gradient descent algorithm called Adam [61]. We also use batch normalization [62] (which has regularization properties) and dropout [63] (which prevents overfitting) to improve the training of the DNN. The problem of determining the Higgs-boson CP state is framed as a binary classification because the aim is to distinguish between the two possible (scalar and pseudoscalar) Higgs CP states.

We consider three separate problems for $H \rightarrow \tau\tau$ channels: $\rho^{\pm} - \rho^{\mp}$, $a_1^{\pm} - \rho^{\mp}$, and $a_1^{\pm} - a_1^{\mp}$. We solve all three problems using the same neural network architecture. Depending on the decay channel for the outgoing τ pairs, each of the cases contains a different number of dimensions to describe an event, i.e., the production of the Higgs boson decaying into a τ -lepton pair. Each data point consists of features which represent the observables/variables of the consecutive event. The data point is thus an event of the Higgs-boson production and decay into a τ -lepton pair. The structure of the event is represented as follows:

$$x_i = (f_{i,1}, \dots, f_{i,D}), \quad w_{a_i}, w_{b_i}. \quad (\text{A1})$$

The $f_{i,1}, \dots, f_{i,D}$ represent numerical features, and w_{a_i} and w_{b_i} are weights proportional to the likelihood that an event comes from a set A or B (binary scalar or pseudoscalar classification). The weights calculated from the quantum field theory matrix elements are available and stored in the simulated data files. This is a convenient situation that does not happen in many other cases of ML classification. The A and B distributions are highly overlapped in $(f_{i,1}, \dots, f_{i,D})$ space; a more detailed discussion can be found in Ref. [20]. Perfect separation is therefore not possible and $w_{a_i}/(w_{a_i} + w_{b_i})$ corresponds to the Bayes optimal probability that an event is sampled from set A and not set B . w_{a_i} and w_{b_i} are used to compute targets during the training procedure.

Because model A and B samples are prepared with the same events and differ only in the spin weights w_{a_i} and w_{b_i} , the statistical fluctuations of the learning procedure are greatly reduced. This also has consequences for the actual implementation of the DNN metric and code.

To quantify the classification performance, weighted AUC and APS scores are used [64], and we have not explored further alternatives. For each data point x_i , the DNN returns a probability p_i that it is correctly (not correctly) classified as being of type A and it contributes to the final loss function twice, with weight w_{a_i} and w_{b_i} , respectively. With this definition, an AUC = 0.5 would be obtained for a random assignment, while an AUC = 1.0 would be obtained for perfect separation. As in the studied problems, the distributions overlap and the best achievable AUC $\simeq 0.78$ is obtained only with $p_i = w_{a_i}/(w_{a_i} + w_{b_i})$ (oracle predictions). The value slightly depends on the case studied, due to the applied minimal set of cuts on

the kinematics of τ decay products to partly emulate detector conditions.

Weighted events with w_a and w_b are used for convenience and to limit statistical fluctuations. We have repeated a part of the DNN classification chain (training, validation, testing) using unweighted events too [66] and have found very good performance consistency.

The DNN architecture [67] consists of D -dimensional input (feature list) followed by six layers of 300 nodes each with ReLU [68] activation functions and a one-dimensional output layer returning the probability p_i of the choice. It is calculated using the softmax function. The metric minimized by the model is the negative log likelihood of the true targets under Bernoulli distributions.

The parameter that was optimized with respect to what was used in Ref. [20] was a dropout [63]. For the analysis presented in Ref. [20], the AUC score was obtained after training with five epochs. It was considered sufficient. Here, given the variety of feature lists, we performed training on a much larger number of epochs and studied the optimal working point for the number of epochs and dropout. The 20% dropout seemed to be the optimal choice to avoid overfitting, which occurs in cases with a larger number of training epochs. The best performance was achieved after 5–50 epochs, depending on the case. Training with 50 epochs was used on the test samples to quote the final AUC score. While optimizing the dropout level, we observed that although it leads to a more robust DNN (smaller risk for overfitting), the performance was sometimes somewhat reduced. The positive impact of the dropout is illustrated (on a data set consisting of one million event samples) in Fig. 8 for the $\rho^\pm - \rho^\mp$ channel and the feature list Variant-1.1 for training and validation.

FIG. 8. The DNN training and validation AUC score as a function of the number of epochs, for the $\rho^\pm - \rho^\mp$ channel and the sample feature list Variant-1.1, without dropout (top) and with dropout = 0.2 (bottom).

APPENDIX B: ALTERNATIVE ML TECHNIQUES

Although deep neural networks are often used for classification tasks in high-energy physics, many other more classical techniques are used as well and are often able to achieve similar classification performance. Despite promising results that are often reported in papers, one should always remember that a machine-learning technique that performs well on one data set with specific features may deliver not so promising results on another.

This observation is of a fundamental nature and results from the mathematical assumption behind particular ML libraries. The solutions that were developed for the libraries depend on the application domains the particular systems were prepared for. Reference [69] can be used as a guide. The recent study in Ref. [70] compared several machine learning algorithms. Despite not being able to contribute much to the topic, we show that for the application discussed throughout this paper it is indeed the case that the DNN technique by far outperforms the more classical approaches.

FIG. 9. The AUC score of the BT classifier as a function of tree depth.

The following machine learning techniques were chosen for the comparative study:

- (1) Boosted trees [38].
- (2) Random forests [39].
- (3) Support vector machines [40].

For BTs the XGBoost [38] library was used, while for SVMs and RFs the scikit-learn library [71] was chosen.

The AUC scores were used to evaluate the performance of the ML methods. This is one of the recommended approaches when using a single number in the evaluation of machine learning algorithms on binary classification problems. To minimize bias the comparisons were carried out on the same data sets [72].

For the BT method the point of interest was to check the dependence of the obtained results on the depth of a tree. The AUC score as a function of tree depth is given in Fig. 9 for the $\rho^\pm - \rho^\mp$ case and several feature lists. As depth affects the complexity of the computation, a search for an optimal choice in terms of both results and efficiency of computation was performed. A tree depth between 3 and 10 was suggested in Ref. [38]. At first we have used depths from 3 to 20. The upper bound was increased to see the trend in the AUC score plots. The results seemed to increase up to 20. Additional evaluations with the depth equal to the number of features of a given Variant- $X.Y$ were also carried out.

As suggested in the literature [73], for the RF method 128 trees (estimators) were used. For the next-best split during the tree building number of features equal to $\log_2(N_f)$ or $\sqrt{N_f}$, where N_f denotes number of original features, was tried. The performances of the two choices are comparable. The optimal tree-depth value found for BTs was used. Tests with a larger number of trees (300) and greater tree depths (30) were also carried out.

TABLE V. AUC scores for different ML methods and different feature sets obtained for the $\rho^\pm - \rho^\mp$ validation sample.

Features set	DNN	BT	RF	SVM
Variant-All	0.764	0.641	0.626	0.635
Variant-1.0	0.655	0.569	0.567	0.559
Variant-2.1	0.657	0.564	0.564	0.572
Variant-2.2	0.657	0.562	0.559	0.568

TABLE VI. Training time, memory usage, and efficiency information as reported by the job submission system. The $\rho^\pm - \rho^\mp$ sample with the feature list Variant-1.0 was used and different ML methods were compared. The set of 800 000 training data points was used for DNN, BT, and RF, while for SVM only 100 000 data points were used.

Method	# Data points	Training time (h:min:sec)	Memory usage (% of 20 GB)	Efficiency (CPU)
DNN	800 000	01:21:58	6.3%	68.8%
SVM	100 000	01:52:03	4.3%	25.0%
BT	800 000	00:09:40	9.6%	24.9%
RF	800 000	00:34:25	37.9%	96.1%

For the SVM method, tests were first performed to determine which kernel—linear or radial basis function (RBF)—gives more promising results. This resulted in choice of RBF kernel with more stable results. In the next step, the parameters C and γ (soft margin and kernel parameter) were fine-tuned. The parameters were evenly distributed on a logarithmic scale from 10^{-3} to 10^3 . To avoid excessive computation, fine-tuning was performed only for Variant-All, and on a smaller event sample. The obtained parameters, $C = 10$ and $\gamma = 0.1$, were then used to train the classifier on other feature lists as well.

A comparison of the best performance for the $\rho^\pm - \rho^\mp$ channel and different feature lists is shown in Table V. Clearly, the performance of the DNN is outstanding.

We compared the execution times, memory usage, and efficiency for all of the classifiers. We used the Prometheus cluster [74]. Jobs were executed at one node with four tasks per node and 5 GB of memory per task. The comparison is reported in Table VI. By far, the SVM training took the longest amount of time. The BT training took the least amount of time—under 10 minutes—which made it 8 times faster to train than the DNN. Both the DNN and RF used the resources well, achieving efficiencies of 68.8 and 96.1%, respectively.

The training times varied significantly between the ML methods. This is probably due to relations between the features that are unexpected by ML. The Higgs CP state classification turned out to be a challenge for more classical ML algorithms.

- [1] D. Guest, K. Cranmer, and D. Whiteson, *Annu. Rev. Nucl. Part. Sci.* **68**, 161 (2018).
- [2] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, *arXiv:1903.10563*.
- [3] K. Albertsson *et al.*, *J. Phys. Conf. Ser.* **1085**, 022008 (2018).
- [4] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, *Phys. Rev. D* **93**, 094034 (2016).
- [5] C. Shimmmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Sogaard, *Phys. Rev. D* **96**, 074034 (2017).
- [6] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban, and D. Whiteson, *Phys. Rev. D* **94**, 112002 (2016).
- [7] G. Louppe, K. Cho, C. Becot, and K. Cranmer, *J. High Energy Phys.* **01** (2019) 057.
- [8] K. Fraser and M. D. Schwartz, *J. High Energy Phys.* **10** (2018) 093.
- [9] G. Kasieczka *et al.*, *SciPost Phys.* **7**, 014 (2019).
- [10] G. Kasieczka, N. Kiefer, T. Plehn, and J. M. Thompson, *SciPost Phys.* **6**, 069 (2019).
- [11] M. Kramer, J. H. Kuhn, M. L. Stong, and P. M. Zerwas, *Z. Phys. C* **64**, 21 (1994).
- [12] G. R. Bower, T. Pierzchala, Z. Was, and M. Worek, *Phys. Lett. B* **543**, 227 (2002).
- [13] A. Rouge, *Phys. Lett. B* **619**, 43 (2005).
- [14] K. Desch, Z. Was, and M. Worek, *Eur. Phys. J. C* **29**, 491 (2003).
- [15] S. Berge and W. Bernreuther, *Phys. Lett. B* **671**, 470 (2009).
- [16] S. Berge, W. Bernreuther, and S. Kirchner, *Phys. Rev. D* **92**, 096012 (2015).
- [17] ATLAS Collaboration, CERN Report No. ATL-PHYS-PUB-2019-008, 2019.
- [18] P. Baldi, P. Sadowski, and D. Whiteson, *Nat. Commun.* **5**, 4308 (2014).
- [19] P. Baldi, P. Sadowski, and D. Whiteson, *Phys. Rev. Lett.* **114**, 111801 (2015).
- [20] R. Jozefowicz, E. Richter-Was, and Z. Was, *Phys. Rev. D* **94**, 093001 (2016).
- [21] E. Barberio, B. Le, E. Richter-Was, Z. Was, D. Zanzi, and J. Zaremba, *Phys. Rev. D* **96**, 073002 (2017).
- [22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2017).
- [23] T. Przedzinski, E. Richter-Was, and Z. Was, *Eur. Phys. J. C* **74**, 3177 (2014).
- [24] For definition see e.g., [25].
- [25] M. Davier, L. Duflot, F. Le Diberder, and A. Roug, *Phys. Lett. B* **306**, 411 (1993).
- [26] J. H. Kuhn, *Phys. Rev. D* **52**, 3128 (1995).
- [27] M. Aaboud *et al.* (ATLAS Collaboration), *Phys. Rev. D* **99**, 072001 (2019).
- [28] S. Chatrchyan *et al.* (CMS Collaboration), *J. High Energy Phys.* **05** (2014) 104.
- [29] G. Aad *et al.* (ATLAS Collaboration), *Eur. Phys. J. C* **76**, 295 (2016).
- [30] A. Sirunyan *et al.* (CMS Collaboration), *J. Instrum.* **13**, P10005 (2018).
- [31] M. Aaboud *et al.* (ATLAS Collaboration), *J. Instrum.* **11**, P04008 (2016).
- [32] A. Sirunyan *et al.* (CMS Collaboration), *J. Instrum.* **13**, P05011 (2018).
- [33] Expected performance of the τ decay vertices measurements based on the collisions data have not been published by LHC experiments yet.
- [34] T. Sstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, *Comput. Phys. Commun.* **191**, 159 (2015).
- [35] N. Davidson, G. Nanava, T. Przedzinski, E. Richter-Was, and Z. Was, *Comput. Phys. Commun.* **183**, 821 (2012).
- [36] This is 10 times larger statistics than used in [20].
- [37] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, *Eur. Phys. J. C* **76**, 235 (2016).
- [38] T. Chen and C. Guestrin, *arXiv:1603.02754*.
- [39] L. Breiman, *Mach. Learn.* **45**, 5 (2001).
- [40] A. Bevan, R. Goni, and T. Stevenson, *J. Phys. Conf. Ser.* **898**, 072021 (2017).
- [41] In the present paper we were able to improve with respect to [20] mainly thanks to 10 times larger training samples.
- [42] A. P. B. Bradley, *Pattern Recognit.* **30**, 1145 (1997).
- [43] T. Fawcett, *Pattern Recogn. Lett.* **27**, 861 (2006).
- [44] Consequence of τ decay dynamic. See e.g., Ref. [26].
- [45] K. Desch, A. Imhof, Z. Was, and M. Worek, *Phys. Lett. B* **579**, 157 (2004).
- [46] This point may become important for discussion of ambiguities due to missing p_T of jets accompanying the Higgs production. Then, it may be helpful to have 3 constraints which may be used e.g., for missing p_T^{miss} generated by jets heavy flavor resonances of decays with neutrinos, contribute to E_T^{miss} as well.
- [47] True or truth level is a short-cut denoting that it is calculated from the generated event kinematic, without any approximation or smearing.
- [48] The Gaussian shape, as can be concluded from central limit theorem, the universal statistical distribution for the variable obtained as an average of large set of independent stochastic variables, may be too simplistic to use as a test example. Also, because the lifetime of the τ follows an exponential distribution we have chosen such a shape for the smearing of ϕ_v^{smear} which can be, in real detector response simulations, proportional to the (inverse) of the distance between τ decay and production vertices. We have used other distributions (like Gaussian distribution with exponentially enhanced tail) as well and conclusions on the DNN algorithm performances remained similar. The DNN could learn from inaccurate distributions. The exponential tail was supposed to introduce a penalty for learning. The actual distribution depends on details of the detectors and reconstruction algorithms used by the experiments. This challenging effort, requiring all details of its geometry, has not been completed so far. Preliminary results are not available publicly, in fact they are of considerable complexity and depend on the detector regions (barrel, endcap etc.) as well as on the τ -lepton energy and decay channel.
- [49] V. Cherepanov and A. Zotz, *arXiv:1805.06988*.
- [50] M. Tanabashi *et al.* (Particle Data Group), *Phys. Rev. D* **98**, 030001 (2018).
- [51] E. Zhang and Y. Zhang, *Encyclopedia of Database Systems* (Springer, New York, 2009), p. 192.
- [52] Because of physics properties, they should be the same for all channels, but as we are filtering events and finite statistics, they differ on third digit.

- [53] Main results of Ref. [20], where only `Variant-1.X` were studied have been recalled in Sec. II, Table I. Nonetheless for overall consistency, we have reevaluated some of those results again.
- [54] D. Jeans, *Nucl. Instrum. Methods Phys. Res., Sect. A* **810**, 51 (2016).
- [55] The sub-sub-index β encodes the size of the smearing parameter β .
- [56] Polynomial modification is implemented in validation and test samples with the help of Monte Carlo unweighting: $wt = f_{\text{valid}}(\Delta\phi_\nu, \beta, b, c) / f_{\text{train}}(\Delta\phi_\nu, \beta)$.
- [57] We hope, that these degradation parameters will be replaced in the future by sophisticated detector simulations. Our evaluation indicates, that already modest and even partial reconstruction of the τ decay vertex position is useful. The experimental effort, like Refs. [29–32] and as mentioned in [58], is encouraging. Our example can not substitute future work with well understood detection details. Nonetheless it hints for a possible method of experimental ambiguities evaluation in DNN applications.
- [58] V. Cherepanov, E. Richter-Was, and Z. Was, *SciPost Phys. Proc.* **1**, 018 (2019).
- [59] M. Erdmann, E. Geiser, Y. Rath, and M. Rieger, *J. Instrum.* **14**, P06006 (2019).
- [60] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, Software available from tensorflow.org (2015), 1.
- [61] D. Kingma and J. Ba, [arXiv:1412.6980](#).
- [62] S. Ioffe and C. Szegedy, [arXiv:1502.03167](#).
- [63] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *J. Mach. Learn. Res.* **15**, 1929 (2014).
- [64] In Ref. [65] we have demonstrated that *for our applications* the AUC score approximate reasonably well the probability to identify single event as scalar or pseudoscalar, thus can be used to calculate statistical significance of the event sample as well.
- [65] P. Bialas, D. Nemeth, and E. Richter-Was, [arXiv:1803.00838](#).
- [66] Randomly sampling events with weights for model A and B. Then only such events were used for training.
- [67] The choice of the DNN architecture was optimised with not presented here studies. A variety of the activation functions, number of layers and number of nodes was tried. The configuration used in [20] was confirmed as optimal one.
- [68] A. F. Agarap, [arXiv:1803.08375](#).
- [69] M. Fernandez-Delgado *et al.*, *J. Mach. Learn. Res.* **15**, 3133 (2014).
- [70] C. Zhang, L. Changchang, Z. Xiangliang, and G. Almpanidis, *Expert Syst. Appl.* **82**, 128 (2017).
- [71] F. Pedregosa *et al.*, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [72] In case of SVM subsets were used due to CPU limitations.
- [73] T. Oshiro, P. Perez, and J. Baranauskas, How many trees in a random forest? in *International Workshop on Machine Learning and Data Mining in Pattern Recognition* (Springer, Berlin, Heidelberg, 2012).
- [74] Prometheus Computing Cluster, Academic Computing Centre CYFRONET, Krakow, Poland.