Interpretable boosted-decision-tree analysis for the MAJORANA DEMONSTRATOR

I. J. Arnquist,[1] F. T. Avignone, III,[2,3] A. S. Barabash,[4] C. J. Barton,[5] K. H. Bhimani,[6,7] E. Blalock,[8,7] B. Bos,[6,7] M. Busch,[9,7] M. Buuck,[10,*] T. S. Caldwell,[6,7] Y-D. Chan,[11] C. D. Christofferson,[12] P.-H. Chu,[13] M. L. Clark,[6,7] C. Cuesta,[14] J. A. Detwiler,[10] Yu. Efremenko,[15,3] S. R. Elliott,[13] G. K. Giovanetti,[16] M. P. Green,[8,7,3] J. Gruszko,[6,7] I. S. Guinn,[6,7] V. E. Guiseppe,[3] C. R. Haufe,[6,7] R. Henning,[6,7] D. Hervas Aguilar,[6,7] E. W. Hoppe,[1] A. Hostiuc,[10] M. F. Kidd,[17] I. Kim,[13] R. T. Kouzes,[1] T. E. Lannen V.,[2] A. Li,[6,7,†] J. M. López-Castaño,[3] E. L. Martin,[6,7,‡] R. D. Martin,[18] R. Massarczyk,[13] S. J. Meijer,[13] T. K. Oli,[5] G. Othman,[6,7,§] L. S. Paudel,[5] W. Pettus,[19,20] A. W. P. Poon,[11] D. C. Radford,[3] A. L. Reine,[6,7] K. Rielage,[13] N. W. Ruof,[10] D. C. Schaper,[13] D. Tedeschi,[2] R. L. Varner,[3] S. Vasilyev,[21] J. F. Wilkerson,[6,7,3] C. Wiseman,[10] W. Xu,[5] and C.-H. Yu[3]

(MAJORANA Collaboration)

[1]*Pacific Northwest National Laboratory, Richland, Washington 99354, USA*
[2]*Department of Physics and Astronomy, University of South Carolina, Columbia, South Carolina 29208, USA*
[3]*Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, USA*
[4]*National Research Center "Kurchatov Institute" Institute for Theoretical and Experimental Physics, Moscow 117218, Russia*
[5]*Department of Physics, University of South Dakota, Vermillion, South Dakota 57069, USA*
[6]*Department of Physics and Astronomy, University of North Carolina, Chapel Hill, North Carolina 27514, USA*
[7]*Triangle Universities Nuclear Laboratory, Durham, North Carolina 27708, USA*
[8]*Department of Physics, North Carolina State University, Raleigh, North Carolina 27695, USA*
[9]*Department of Physics, Duke University, Durham, North Carolina 27708, USA*
[10]*Center for Experimental Nuclear Physics and Astrophysics, and Department of Physics, University of Washington, Seattle, Washington 98195, USA*
[11]*Nuclear Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[12]*South Dakota Mines, Rapid City, South Dakota 57701, USA*
[13]*Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[14]*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, CIEMAT, 28040 Madrid, Spain*
[15]*Department of Physics and Astronomy, University of Tennessee, Knoxville, Tennessee 37916, USA*
[16]*Physics Department, Williams College, Williamstown, Massachusetts 01267, USA*
[17]*Tennessee Tech University, Cookeville, Tennessee 38505, USA*
[18]*Department of Physics, Engineering Physics and Astronomy, Queen's University, Kingston, Ontario K7L 3N6, Canada*
[19]*Department of Physics, Indiana University, Bloomington, Indiana 47405, USA*
[20]*IU Center for Exploration of Energy and Matter, Bloomington, Indiana 47408, USA*
[21]*Joint Institute for Nuclear Research, Dubna 141980, Russia*

(Received 22 July 2022; accepted 15 November 2022; published 25 January 2023)

The MAJORANA DEMONSTRATOR is a leading experiment searching for neutrinoless double-beta decay with high purity germanium (HPGe) detectors. Machine learning provides a new way to maximize the amount of information provided by these detectors, but the data-driven nature makes it less interpretable compared to traditional analysis. An interpretability study reveals the machine's decision-making logic, allowing us to learn from the machine to feed back to the traditional analysis. In this work, we present the first machine learning analysis of the data from the MAJORANA DEMONSTRATOR; this is also the first interpretable machine learning analysis of any germanium detector experiment. Two gradient boosted decision tree models are trained to learn from the data, and a game-theory-based model interpretability study is conducted to understand the origin of the classification power. By learning from data, this analysis recognizes the correlations among reconstruction parameters to further enhance the background rejection performance. By learning from the machine, this analysis reveals the importance of new background categories to reciprocally benefit the standard MAJORANA analysis.

[*]Present address: SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA.
[†]Corresponding author: liaobo77@ad.unc.edu
[‡]Present address: Duke University, Durham, NC 27708, USA.
[§]Present address: Universität Hamburg, Institut für Experimentalphysik, Hamburg, Germany.

This model is highly compatible with next-generation germanium detector experiments like LEGEND since it can be simultaneously trained on a large number of detectors.

DOI: 10.1103/PhysRevC.107.014321

I. INTRODUCTION

Neutrinoless double beta decay ($0\nu\beta\beta$) [1–3] is a hypothetical lepton number violating process ($\Delta L = 2$) beyond the standard model. The observation of $0\nu\beta\beta$ would prove that the neutrino is its own antiparticle, also known as the Majorana particle. This is a key ingredient for leptogenesis [4], which is one model that explains the observed matter-antimatter asymmetry in our universe. Measuring $0\nu\beta\beta$ is a challenging task since it occurs with an ultralong half-life ($>10^{26}$ yr) [5,6]. This limits the number of signal events we can collect, and requires us to reliably discover them among a plethora of backgrounds. To maximize their discovery potential, germanium-based $0\nu\beta\beta$ searches seek to operate in the quasi-background-free regime, where less than one background event is expected in the region of interest over the full lifetime of the experiment. Therefore, the ability to suppress background as much as possible is pivotal to $0\nu\beta\beta$ search experiments.

Traditional background suppression techniques are typically derived from physical first principles, which are used to define event-level reconstruction parameters. A cut is then placed upon the reconstruction parameters to minimize backgrounds while retaining signals. Because a traditional analysis begins with first principles, interpretability is inherently built into this approach. However, there are weaknesses with this approach as well. The actual response of a detector to a particular background source is often clouded by complex effects inherent to the detector technology that are difficult to model, reducing the effectiveness of any background rejection cuts. Furthermore, many physical effects that produce backgrounds must be handled individually, increasing the chances that a particular source of background will be neglected. Finally, unknown detector physics could also produce potential bias in reconstruction parameters, harming the performance of traditional background cuts.

Machine learning presents an alternative to the traditional first-principles approach to background rejection, and has already been proven quite successful for neutrino physics experiments [7–15]. Unlike traditional analyses, the background suppression power of machine learning algorithms comes primarily from data. This allows machine learning models to efficiently handle unknown backgrounds to reach state-of-the-art performance. Unfortunately, learning from data makes machine learning analyses less interpretable compared to the traditional ones. Therefore, many machine learning analyses are equipped with an interpretability study to reveal the underlying decision-making logics [16–18].

In this work, we present the first machine learning analysis for the MAJORANA DEMONSTRATOR [19–21], which is also the first interpretable machine learning analysis of any germanium detector experiment. This analysis was inspired by the drift-time correction to our multisite and surface alpha discrimination parameters, which indicated that accounting for correlations between parameters could enhance background suppression power. We constructed two boosted decision tree (BDT) models to reject two of the most critical backgrounds in the MAJORANA DEMONSTRATOR, namely the MSBDT for multisite events and the αBDT for alpha events. Both models take individual reconstruction parameters as inputs and are trained on the detector data to provide background suppression. By learning from the data, this model utilizes multivariate correlations among reconstruction parameters to improve the background suppression. It also reduces the need for detector- and run-level tuning, which would be time consuming in future large-scale experiments such as LEGEND [22].

In addition, we conducted a comprehensive interpretability study to understand the source of classification power. This study leverages a coalitional game theory concept to unravel the black box that is the inside of a machine learning model [23]. It has been widely used in biomedical science [24–26] and other fields [27–29]. By learning from the machine, we verified our BDTs' abilities to learn multivariate correlations among different features. Furthermore, we revealed the importance of new background categories that the traditional, first-principles-based analysis did not address, which eventually led to new analysis cuts in the standard MAJORANA analysis.

The paper is structured as follows. Section II describes the MAJORANA DEMONSTRATOR experiment, the major background sources, and the traditional analysis cuts to reject them. Section III describes the data pipeline for collecting and preprocessing the training data. Section IV describes the gradient BDT algorithms. Section V reports the training results of the MSBDT and the αBDT with a comparison to the standard MAJORANA analysis. Section VI describes the interpretability study we conducted. We highlight Sec. VI B, which outlines the ability of machine learning to reveal the importance of new background categories and reciprocally benefit the standard MAJORANA analysis pipeline.

II. MAJORANA DEMONSTRATOR

The MAJORANA DEMONSTRATOR experiment searches for $0\nu\beta\beta$ decay in ^{76}Ge using 40.4 kg of high purity germanium (HPGe) detectors [19]. Of these, 27.2 kg of p-type point-contact (PPC) HPGe detectors are enriched to 88% in ^{76}Ge [30]. The DEMONSTRATOR is operated at the 4850-ft level of the Sanford Underground Research Facility in Lead, South Dakota. Data were taken from August 2015 to March 2021, and are split into nine data set (DS) periods, referred to as DS0–DS8. Starting with DS8 (August 2020), novel p-type inverted-coaxial point-contact (ICPC) detectors [31] were added to the MAJORANA DEMONSTRATOR detector array. Data taking finished in March 2021, with a total enriched

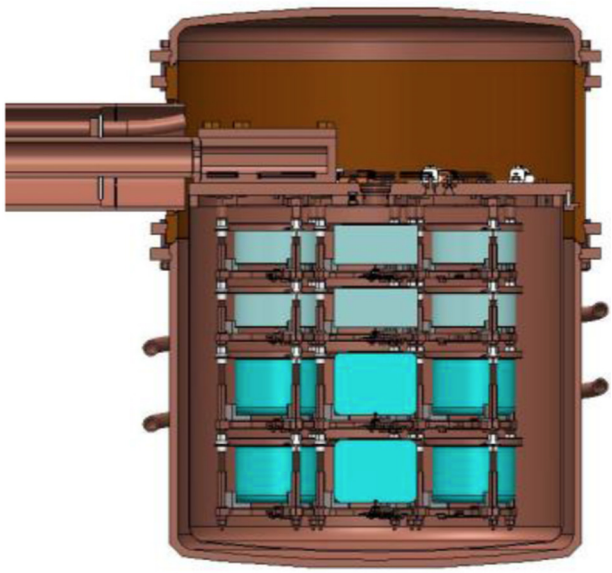


FIG. 1. A diagram of one MAJORANA DEMONSTRATOR detector module and the HPGe detectors within.

exposure of 64.5 kg yr, 2.82 kg yr of which is from ICPC detectors [32]. The DEMONSTRATOR's HPGe detectors, in combination with low-noise electronics [33], have achieved good linearity over a broad energy range [34], and best-in-field energy resolution with a full width at half maximum (FWHM) approaching 0.1% at the $Q_{\beta\beta}$ (2039 keV) of ^{76}Ge [21]. This excellent energy performance, coupled with the low energy threshold and low background of the DEMONSTRATOR, makes it a competitive $0\nu\beta\beta$ decay experiment. One module of the DEMONSTRATOR is shown in Fig. 1.

A weekly calibration is conducted to monitor detector stability and provide data for developing analysis cuts. The thorium isotope ^{228}Th was selected as the primary calibration source because its decay chain emits several gamma rays spanning from a few hundred to 2615 keV, which covers the $Q_{\beta\beta}$ of ^{76}Ge and allows for calibration over a wide energy range. During calibrations, the ^{228}Th source is deployed into the calibration track, which surrounds the cryostat in a helical path [35]. Event energies are tuned to minimize ^{228}Th calibration source gamma line width [36]. This routine calibration provides an excellent source of training data that will be discussed in Sec. III.

Most $0\nu\beta\beta$ events are single-site events which deposit all of their charge in a single location of <1 mm linear dimension in a detector. This type of event appears in a MAJORANA DEMONSTRATOR detector as a waveform with a single sharply rising step, as indicated by the black trace in the top panel of Fig. 2(a). Since the waveform itself is the integrated ionized charge collected from an energy deposition in the detector, the derivative of the waveform (red traces in the same panel) is effectively the current induced as charges drift towards the point contact. In the MAJORANA DEMONSTRATOR, two major background sources are multisite events and surface-alpha events. If charge is deposited at multiple locations within the crystal, the drift times may differ up to ≈1 μs, resulting in a waveform with multiple steps as shown in the bottom panel of Fig. 2(a). This leads to a current pulse with a smaller maximum value than that of a single-site event with the same energy. Based on this first principle, we designed the current amplitude vs energy (`AvsE`) described in Ref. [37]. The current amplitude is estimated by a linear fit to a smaller range of the waveform. Cutting on the energy-normalized current amplitude (or `AvsE`) leads to efficient multisite event rejection. In the standard MAJORANA analysis, we select events from a dedicated `AvsE` range by applying both low and high `AvsE` cuts.

The other major source of backgrounds at $Q_{\beta\beta}$ is from the alpha particles impacting the passivated surface and p+ contact surfaces of detectors. Prior to the most recent MAJORANA DEMONSTRATOR data release [32], this background source was rejected entirely using the first principle of the "delayed charge recovery" effect [38]. Based on the characteristics of alpha interactions, it appears that charge mobility is drastically reduced on or near the passivated surface. Therefore, a fraction of the charge from these interactions is slowly released on the timescale of waveform digitization, leading to

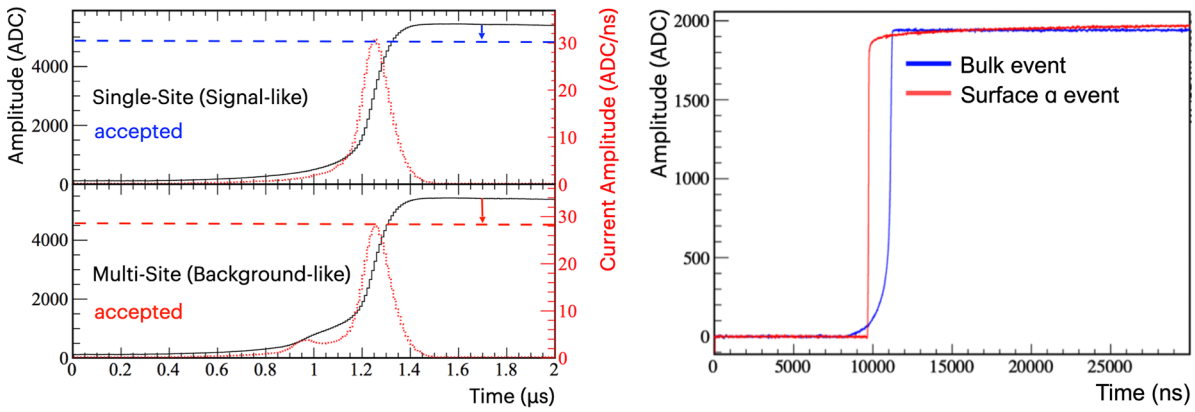


FIG. 2. (a) Pulse shape plot of single-site events (top) and multisite events (bottom). The black line shows the raw waveform in ADC (analog-to-digital converter) counts and the red line shows the waveform in current amplitude. (b) Illustration of waveforms from surface alpha event and bulk event.

a measurable increase in the slope of the waveform tail. The delayed charge recovery effect lowers the peak amplitude and, as shown in Fig. 2(b), results in a slowly rising tail slope that distinguishes this event from a bulk event at the same energy. The delayed charge recovery (DCR) cut tags events with larger tail slopes to efficiently reject surface alphas.

At a later stage of the standard MAJORANA analysis, a novel analysis cut based on the first principles of late charge (LQ) was developed. The LQ cut probes the top of the rising edge of the waveform to identify delayed charge collection on $\approx$1 μs timescale. It efficiently eliminates events in the transition layer and an additional population of near-point-contact events.

The final MAJORANA DEMONSTRATOR standard analysis is described in Ref. [32]. The standard MAJORANA analysis for the PPC detectors is developed with the Germanium Analysis Toolkit (GAT), and is thus referred to as the "GAT analysis." The standard MAJORANA analysis for the ICPC detectors is developed independently from GAT, and is referred to as the "ORNL analysis." Both the GAT and ORNL analyses contain independently developed pulse shape discrimination (PSD) cuts derived from the first principles of HPGe detector charge collection: current amplitude versus energy, the delayed charge recovery effect, as well as the late charge effect. In this paper, we denote those cuts as GAT AvsE/DCR/LQ cut and ORNL AvsE/DCR/LQ cut, respectively.

During the development of the standard MAJORANA analysis, we observed that the PSD parameters vary with the length of time it takes the charges to drift to the p+ electrode, defined as drift time, due to well-understood charge cloud diffusion and bulk charge trapping effects. Therefore, a drift-time correction was made to correct for this correlation. In the following text, we will use the terms "standard AvsE/DCR/LQ" to refer to the GAT analysis for PPC detectors and ORNL analysis for ICPC detectors, respectively. Additionally, we extracted the raw AvsE and raw DCR parameters, which are preliminary versions of standard AvsE/DCR thereby not directly used by the standard MAJORANA analysis. The raw parameters are generated under the GAT framework with a detector- and run-wise energy calibration applied. However, the raw parameters are not drift-time corrected, thus underperforming compared to the standard MAJORANA analysis parameters. In this work, we decided to use the raw AvsE/DCR parameters to train the BDTs, and then compare the training results to the standard MAJORANA analysis parameters. The LQ parameters are introduced at a later stage of the standard MAJORANA analysis, thus we decided not to incorporate them to train the machine learning analysis. However, we did include LQ when comparing the two analyses in Sec. V C.

III. DATA PIPELINE

We collected both a signal and a background dataset to train the BDT. The signal dataset should be representative of the signal ($0\nu\beta\beta$ events in our case), and the background dataset should represent the proper background to reject. Before creating signal and background datasets from the MAJORANA data, a standard suite of cuts is applied: periods of high noise associated with liquid nitrogen fills or unstable operation are removed; nonphysical waveforms, pileup waveforms, and pulser events are then removed by data cleaning cuts; and finally events in which multiple germanium detectors are triggered are removed. We particularly avoided the usage of high-level selection cuts, such as standard AvsE/DCR/LQ, since the tuning and validation of these cuts can be time consuming. Decoupling from these cuts allows a fast-track application of this model on newly taken data from multiple detectors. We then chose events in the double escape peak (DEP) from ^{228}Th calibration data as the signal dataset. The DEP events are pair production events where both gammas have successfully escaped from the detector, thus they are mostly single-site events. An energy cut of 1592.5 $\pm$ 2.5 keV is applied to select DEP events. Monte Carlo simulations including x-ray excitations and bremsstrahlung predict the events under DEP selection criteria to be 90% single site with 10% multisite impurities. The background datasets for the MSBDT and the αBDT are selected separately. For MSBDT, we select events under the single escape peak (SEP) of ^{228}Th calibration data. The SEP events are pair production events where only one gamma has escaped from the detector. Although all SEP events are technically multisite, if those sites occur at the isochrones of equal drift time, they will reach the point contact at roughly the same time. In that case, the time difference between the two sites is smaller than the detector's timing resolution, resulting in an apparently single-site waveform even though more than one energy deposition has occurred. These events are the impurities to the background dataset. An energy cut of 2103.5 $\pm$ 2.5 keV is applied to select these events.

For the αBDT, since the major alpha backgrounds are energy-degraded alpha events from the continuum, we have to select training data from a broader spectrum. We first select high energy alpha events by collecting events above 2615 keV in background runs. This sample is expected to have some contamination from high energy gamma events, originating from the decay of cosmogenically- and neutron-induced isotopes and uranium/thorium chain. We then select low energy alpha events by collecting the DCR tagged background events in a 1000–2615 keV energy range. In this way, a total of 723 high energy alpha and 2839 low energy alpha events are selected.

After event selection, we extract eight features from every event. The names and descriptions of these parameters are listed in Table I. AvsE and DCR are dedicated pulse shape parameters for multisite and alpha rejection respectively. Other parameters are added to probe their multivariate correlations with AvsE/DCR and to each other. For example, adding the channel parameter will allow the model to perform detector-wise tuning, adding tDrift and tDrift50 will allow the model to perform a drift-time correction, and our noise parameter allows us to look for correlations during noisy periods in the data. Among all features in Table I, some features are continuous and some features are categorical. BDTs naturally handle both types of feature in the structure of the tree, allowing us to train on all detectors from all run periods simultaneously.

TABLE I. List of input features to the MSBDT and the αBDT. Cat. stands for categorical and Cont. stands for continuous.

Features	Type	Description
`detType`	Cat.	detector type: enriched PPC or ICPC
`channel`	Cat.	detector DAQ channel
`tDrift`	Cont.	drift time from the start of the rise to 99% waveform amplitude
`tDrift50`	Cont.	drift time from the start of the rise to 50% waveform amplitude
`AvsE`	Cont.	raw A vs E, pulse shape parameter for multisite event rejection [32,37]
`DCR`	Cont.	raw DCR, pulse shape parameter for alpha event rejection [32,38]
`noise`	Cont.	measuring 10–20 MHz noise
`DS`	Cat.	data period the event belongs to, defined by run ranges

A. Data augmentation

The data we selected above are highly imbalanced. First of all, only about 4% of data points are provided by ICPC detectors while the rest are from PPC detectors. Secondly, only 3562 α events are collected compared to 600 000 DEP events in αBDT training. This forms a typical long tail distribution where the head class contains most of the events and the tail class contains only a minimal proportion. If a BDT is trained with such an imbalanced dataset, it will be heavily biased towards the ample head class while ignoring the scarce tail class. We fix this issue by performing data augmentation.

Data augmentation refers to algorithms that generate synthetic data points to boost the population of the tail class for training purposes. We employ it to boost the population of both ICPC detector events and surface alpha events. The input dataset contains eight features per event, three of which are categorical features. A synthetic minority oversampling technique–nominal and continuous (SMOTE-NC) [39] algorithm is adopted for data augmentation. SMOTE-NC generates synthesized data by randomly interpolating between data points and their nearest neighbors. It works well on low dimensional data with both continuous and categorical features. It is first applied to all three datasets (DEP, SEP, and alpha) to boost the population of ICPC events by a factor of $\approx$50, then applied again on the alpha dataset to boost the population of alpha events by a factor of $\approx$115. We refer to the events directly collected from detectors as genuine events and the events from data augmentation as augmented events. The augmented events are only used for training; model evaluation that will be discussed in Sec. V is based on genuine events.

B. Distribution matching

While building our BDT models, we want them to look at the correlations of features instead of single features, unless that single feature is the first-principle feature as discussed in Sec. II. The first-principle feature—that is `AvsE` for MSBDT or `DCR` for αBDT—is designed to fulfill the same background rejection goal as the BDT model. We do not expect features other than the first-principle feature to contribute to the classification independently, but they can contribute through their correlations with the first-principle feature or other features. Undesirable behavior arises when other parameters are allowed to contribute independently to classification. For example, if a given `channel` in the MSBDT training dataset is accidentally biased to contain 50% more signals than backgrounds, the BDT will "remember" this bias and tend to classify events in this `channel` as signal regardless of the rest of the features. If we then validate the BDT on another out-of-sample, unbiased dataset, the classification performance on this `channel` will be suppressed. This phenomenon is referred to as overfitting. To avoid this kind of overfitting, we carried out a process called "distribution matching" on six out of eight secondary features. Figure 3 shows the distribution matching effect on the `tDrift` feature. The distribution of each secondary feature is first put into a histogram with predefined bin width. Then for every bin in the histogram, we randomly sample without replacement the same number of events from the signal and background datasets. This will reduce the size of both signal and background datasets to the same amount. Eventually, the sampled events are aggregated into a new signal/background dataset. We leave the first-principle feature—`AvsE` for MSBDT and `DCR` for αBDT—unmatched because we expect them to follow different distributions between the signal and background dataset. The `detType` feature is not matched either since it overlaps with the `channel` feature. After distribution matching, the signal/background dataset spectrum will exhibit the same spectrum shape over matched secondary features. Distribution matching only affects the training dataset. The performance of trained BDT is evaluated on both DEP and SEP datasets and a flat Compton continuum (CC) dataset, as described in Sec. V C and Table II.

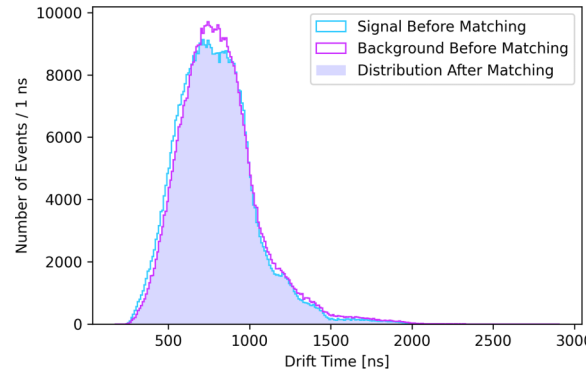


FIG. 3. Distribution matching of the `tDrift` feature in input data.

In summary, the data pipeline for the MSBDT contains the following steps: we first select ^{228}Th DEP events as the signal dataset and ^{228}Th SEP events as the background dataset; we then extract eight features as described in Table I for every event in both datasets; we then perform data augmentation to generate augmented ICPC events; lastly, we perform distribution matching on all features except `AvsE` and `detType`. The data pipeline of the αBDT contains the following steps: we first select ^{228}Th DEP events as the signal dataset, and aggregate both low- and high-energy alphas to form the genuine

TABLE II. Table of survival fractions of signal and background events in each MAJORANA DEMONSTRATOR dataset. The signal and background event selection is defined in Sec. V A for MSBDT and Sec. V B for the αBDT. The BDT cutting thresholds are selected so that they produce the same signal acceptance with the standard analyses. The AvsE corrected, DCR corrected, and LQ parameters are adopted in the standard MAJORANA analysis. CC stands for the Compton continuum (1989–2089 keV) events in the ^{228}Th calibration dataset, and BEW stands for the $0\nu\beta\beta$ background estimation window events (1950–2350 keV, excluding three gamma peaks) in the $0\nu\beta\beta$ search dataset. The survival numbers in BEW are calculated after joint cuts. The number in parentheses (last column of row 11) is the survival number without the LQ cut.

Row Index	Dataset Detector type	DS0 PPC	DS1 PPC	DS2 PPC	DS3 PPC	DS4 PPC	DS5 PPC	DS5c PPC	DS6 PPC	DS6c PPC	DS7 PPC	DS8 PPC	DS8 ICPC	All DS (Expo. weighted)
1	Exposure (kg yr)	1.13	2.24	1.13	0.96	0.26	4.49	2.34	24.52	13.25	4.44	6.41	2.74	64.5
2	Single-site signal (%)	90.3	89.8	88.6	89.9	89.4	89.1	87.4	89.4	89.8	89.7	89.7	88.7	89.5
3	MSBDT bkg. (%)	5.62	5.85	5.01	5.71	6.31	5.95	5.70	5.73	5.58	4.57	6.41	5.76	5.71
4	Standard AvsE bkg. (%)	6.13	6.29	5.93	5.31	5.48	6.24	6.51	6.25	6.39	6.00	6.78	6.17	6.25
5	MSBDT calc. CC (%)	41.9	38.9	36.8	39.7	42.1	41.9	42.6	42.1	40.5	35.6	32.5	31.4	40.3
6	Standard AvsE calc. CC (%)	43.1	42.9	41.5	41.0	42.1	42.0	41.6	42.3	42.7	42.1	43.3	35.0	42.3
7	Bulk event signal (%)	97.9	97.8	98.1	98.9	98.0	97.8	97.8	98.5	98.3	98.5	98.6	97.6	98.2
8	αBDT bkg. (%)	0.4	1.4	2.1	0.8	1.2	3.8	3.4	1.6	1.9	3.5	4.7	8.1	2.1
9	Standard DCR Bkg. (%)	1.7	1.9	2.8	3.9	0.0	2.9	5.7	2.6	3.1	5.4	2.8	0.8	2.9
10	BDT $0\nu\beta\beta$ BEW (#)	11	6	2	0	0	8	6	66	23	20	19	3	164
11	Standard $0\nu\beta\beta$ BEW (#)	11	4	1	0	0	9	5	58	20	17	24	4	153 (168)

alpha dataset; we then extract eight features; perform data augmentation to generate augmented ICPC detector events and alpha events as the background dataset; lastly, we perform distribution matching on all features except DCR and detType.

IV. BOOSTED DECISION TREE

The decision tree (DT) model produces classification decisions by making a series of binary choices. This features allow the decision tree to naturally handle both continuous and categorical dataset, without the need of additional structures such as one-hot encoding. Boosting algorithms allow the machine to generate many decision trees iteratively to form a classification "committee." After training the mth decision tree, the classification committee containing the first through mth trees is denoted $T_m(x_i)$. The dataset can be described as $\{x_i, y_i\}_{i=0}^k$, where x_i is the input event, y_i is the label and k is the number of events. The dataset is modified according to the output of $T_m(x_i)$. The modified dataset is then fed into $T_{m+1}(x_i)$ for training. The way the dataset is modified for each iteration defines the boosting algorithm type. In this work, the BDT model is trained using the LightGBM package [40]. LightGBM adopts a gradient boosting algorithm [41] to grow decision trees. First, a binary cross-entropy loss function $L(y_i, T_m(x_i))$ is defined for the classification task, where y_i is the event label. Then, for each data point, we calculate the pseudo-residual r_{im}:

$$r_{im} = -\frac{\partial L(y_i, T_m(x_i))}{\partial T_m(x_i)}. \tag{1}$$

r_{im} is the negative gradient of the loss function with respect to the classification committee output at x_i. For each boosting iteration, the dataset is modified from $\{x_i, y_i\}_{i=0}^k$ to $\{x_i, r_{im}\}_{i=0}^k$, then a new decision tree $h_{m+1}(x)$ is fit to the modified dataset. The new decision tree is incorporated into the committee via the equation

$$T_{m+1}(x) = T_m(x) + \gamma_{m+1} h_{m+1}(x), \tag{2}$$

where γ_{m+1} is chosen to minimize the loss function by solving the optimization problem

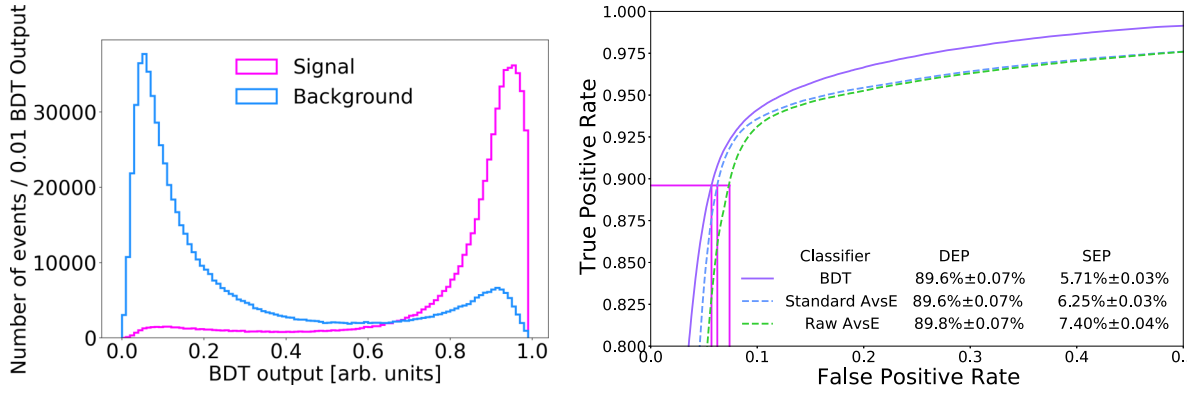
$$\gamma_{m+1} = \arg\min_{\gamma} \sum_{i=0}^{k} L(y_i, T_m(x_i) + \gamma h_{m+1}(x_i)) \tag{3}$$

The procedure above describes the mathematical formulation of BDT training. To train the BDT model, we first mix and shuffle the signal and background datasets. We then split the mixed dataset into training and validation datasets with an 80:20 ratio. The BDT models are trained on the training dataset. An early stopping algorithm will terminate the training process if the loss on the validation dataset does not decrease for a given number of iterations. The performance is quantified on a dedicated evaluation dataset, which will be discussed later in Sec. V; the interpretability study is conducted on a customized interpretability dataset, which will be discussed later in Sec. VI.

LightGBM contains several highly customizable BDT models defined by collections of hyperparameters. Hyperparameters refer to parameters that do not change during training, such as the type of boosting algorithms, maximum number of trees to grow, number of early stopping iterations, and maximum number of leaves per tree. Some hyperparameters may greatly impact the metric, while other parameters may have minimal to no impact. All hyperparameters are searched simultaneously using Bayesian optimization to maximize the background rejection efficiencies at 90% signal acceptance [42].

V. RESULT

After training, we evaluate the performance of both the MSBDT and the αBDT. The trained BDT model takes the



(a) (b)

FIG. 4. (a) MSBDT score distribution for signal and background events. (b) Background subtracted ROC curve for MSBDT classifier, AvsE corrected classifier, and AvsE classifier. The ROC curve plots the true positive rate (TPR) vs the false positive rate (FPR) of a binary classifier by placing the cutting threshold at every possible location. The larger area under ROC curve represents better classification performance. For both AvsE classifiers, only the traditional low AvsE cuts are applied.

input of eight features as described in Table I, and outputs a single floating point number as the classification score between 0 and 1. A higher classification score indicates the input event is more signal-like and a lower classification score indicates the input event is more background-like. A cutting threshold is placed to accept signals and reject backgrounds. The selection criteria of the cutting threshold will be discussed in the following subsections. We also use the receiver operating characteristic (ROC) [43] curve to gauge the classification performance of our models. The ROC curve plots the true positive rate (TPR) vs the false positive rate (FPR) by placing the cutting threshold at every possible location. The fraction of area under the ROC curve (AUC) statistically describes the classification power of a binary classifier, in that larger AUC corresponds to better classification performance and smaller AUC corresponds to worse classification performance. For example, an AUC of 1 indicates perfect classification, and an AUC of 0.5 indicates no classification.

A. MSBDT result

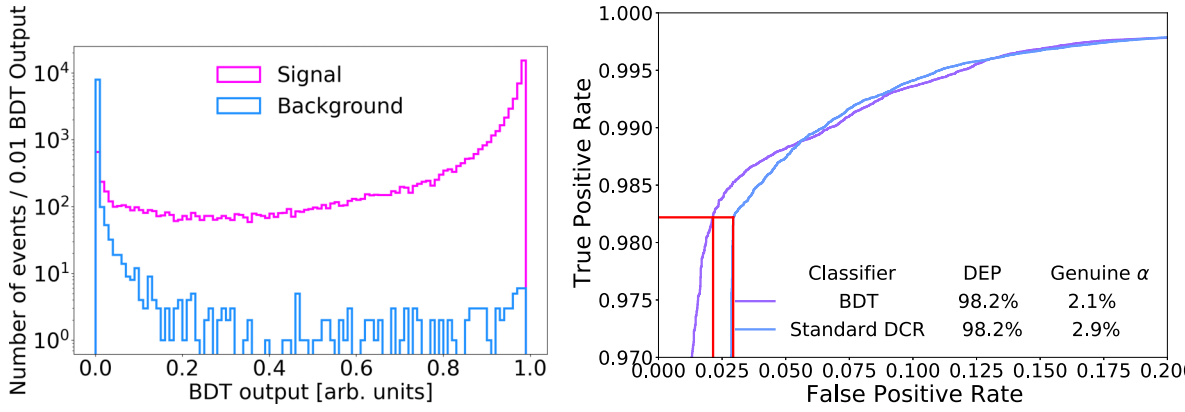
Events from the ^{228}Th calibration data sets are used to test the MSBDT. The DEP events from 1590 to 1595 keV are used as the signal event sample, and the SEP events from 2101 to 2106 keV are used as the background event sample. The MSBDT score spectra for signal and background samples are illustrated in Fig. 4(a). The signal and background peaks are well separated, albeit with "misclassified" events under both peaks. Since both the signal and the background samples have impurities described in Sec. III, these events could be the impure events and thus be correctly classified.

The ROC curves of MSBDT, standard AvsE, and raw AvsE are shown in Fig. 4(b). At each cutting location, the baseline subtraction and uncertainty evaluation are then performed in the same way as in Ref. [37]. To set the MSBDT cutting threshold, we first apply the standard AvsE cut, that is, AvsE > -1.0, to the evaluation dataset. This cut leads to a TPR of 89.6%, shown as the horizontal magenta line in Fig. 4(b). Next, the BDT cutting threshold and raw AvsE cutting threshold are selected to reach the same TPR as the standard AvsE. At this level of TPR, the survival fraction of background samples of the raw AvsE, the standard AvsE, and MSBDT are 7.40%, 6.25%, and 5.71%, respectively. The standard AvsE leverages the drift-time correlations to reject 16.3% of SEP events that the raw AvsE accepts. MSBDT leverages additional multivariate correlations to reject a further 8.6% of SEP events that the standard AvsE accepts.

Rows 2–4 in Table II compare the performance of the MSBDT and the standard AvsE for each MAJORANA dataset. For most datasets, the MSBDT outperforms the standard AvsE on selected data samples. This demonstrates the ability of our BDTs to self-discover the drift-time corrections and other possible feature correlations to improve background rejection performance. Meanwhile, introducing DS, channel, and detType as categorical features allows the machine to perform detector- and run-level tuning without explicitly programming. However, the background data samples described above are only good representations of the true background dataset; the deviation from true background dataset could come from energy (DEP energy vs $Q_{\beta\beta}$ energy) and subtle differences in the intradetector distribution of event positions. Therefore, additional data samples are collected to examine model performance near the true energy region of interest of $0\nu\beta\beta$ decay. These data samples, denoted as calibration Compton continuum (Cal. CC) samples, contain all events between 1989 and 2089 keV from the ^{228}Th calibration runs. Only 40.3% of Cal. CC samples survive MSBDT while 42.3% survive standard AvsE, as shown in rows 5 and 6, Table II.

B. αBDT result

Events from the ^{228}Th calibration data sets and the $0\nu\beta\beta$ search data sets are used to test the αBDT. All ^{228}Th calibration events between 1000 and 2380 keV that pass the standard AvsE cut are selected as the signal samples, and the collection of genuine alpha events are selected as the background samples. The αBDT output distribution of signal and background datasets are shown in Fig. 5(a). Based on the



(a) (b)

FIG. 5. (a) αBDT output distribution for signal and background events. (b) ROC curve for αBDT classifier and standard DCR classifier.

plot, the background dataset is highly concentrated near 0.0 BDT score, indicating an excellent alpha tagging efficiency of the αBDT. Meanwhile, the signal dataset spans the entire range, but most events are still concentrated near 1.0 αBDT output.

The ROC curves of the αBDT and the standard DCR parameter are shown in Fig. 5(b). A cutting threshold is set at the horizontal red line to accept 98.2% of signal events. This acceptance matches the standard DCR acceptance in the standard MAJORANA analysis. At this cutting threshold, the DCR corrected analysis has 2.9% background acceptance, while the αBDT only accepts 2.1% of surface alpha events. As mentioned in Sec. III, low energy genuine alphas are DCR tagged background events between 1000 and 2615 keV. Therefore, when evaluating the performance of the standard DCR cut, 100% of low energy genuine alphas will be manifestly removed. Given the fact that standard DCR is "cheating" on low energy alpha rejection, the αBDT still outperforms standard DCR by rejecting 27.6% of genuine alpha events that standard DCR accepts.

C. Comparison to standard MAJORANA analysis

The background index of the standard MAJORANA analysis is evaluated in the $0\nu\beta\beta$ background estimation window ($0\nu\beta\beta$ BEW). The $0\nu\beta\beta$ BEW samples are collected from 1950 to 2350 keV, excluding ± 5 keV region around the 2039 keV $Q_{\beta\beta}$ value and three gamma peaks at 2103, 2118, and 2204 keV. Based on simulations, the background rate is expected to be flat after the exclusion. Both BDTs are applied to produce a number of survival events for $0\nu\beta\beta$ BEW in each dataset, which can be compared to the number of survival events for each dataset after applying the suites of standard MAJORANA analysis cuts: the standard AvsE cut, the high AvsE cut, the DCR cut, and the LQ cut. Note that neither the LQ feature nor the transition layer events are included in the BDT training process; thus, the BDT analysis will not be sensitive to these events. Given this "unfair" condition, the BDT analysis still manages to match and, in some datasets, outperform the standard MAJORANA analysis. The total number of $0\nu\beta\beta$ BEW survival events for BDT and standard MAJORANA analyses are 164 and 153 [32], indicating consistency between the two. As a comparison, standard MAJORANA analysis without LQ cut allows 168 events to remain in the $0\nu\beta\beta$ BEW. Figure 6 shows the comparison between the two analyses over the entire energy range. The two analyses agrees well except in the low energy region, where the standard MAJORANA analysis cuts more aggressively. This discrepancy is mainly subject to the LQ cut, and the applicability of LQ at low energy is still under investigation. Therefore, these agreements show that the BDT analysis can start from raw parameters and tune them to match a highly optimized analysis.

D. ICPC detectors

For the ICPC detectors, the trained BDT model takes the raw parameters as input and compares the output BDT score to the ORNL analysis parameters. This leads to additional challenges since the raw parameters are developed with GAT while the ORNL parameters are independently developed and customized for ICPC detectors. In this analysis, we use the raw parameters as inputs to train the BDT models to reach or exceed the background rejection performance of the ORNL analysis. This means that the BDT must perform multivariate corrections and account for the technical differences between two independently developed analyses. The second-to-last column of Table II shows the model evaluation results on ICPC detectors. MSBDT outperforms the ORNL AvsE for multisite event rejection, with a background survival fraction of only 5.76%, compared to 6.17% for the ORNL AvsE. It also makes a significant improvement over its primary input parameter, raw AvsE, which has a 25% background survival fraction (not shown in Table II). On the other hand, the αBDT underperforms the ORNL DCR with a 8.1% genuine alpha survival fraction, compared to 0.8% for the ORNL DCR. However, the αBDT still makes a significant improvement over its primary input parameter, raw DCR, which allows 18% of genuine alphas to survive (not shown in Table II). The BDT analyses account for the technical differences among independently developed analyses to simultaneously analyze different types of germanium detectors. Finally, the results of the BDT analyses indicate that the GAT analysis has the potential to

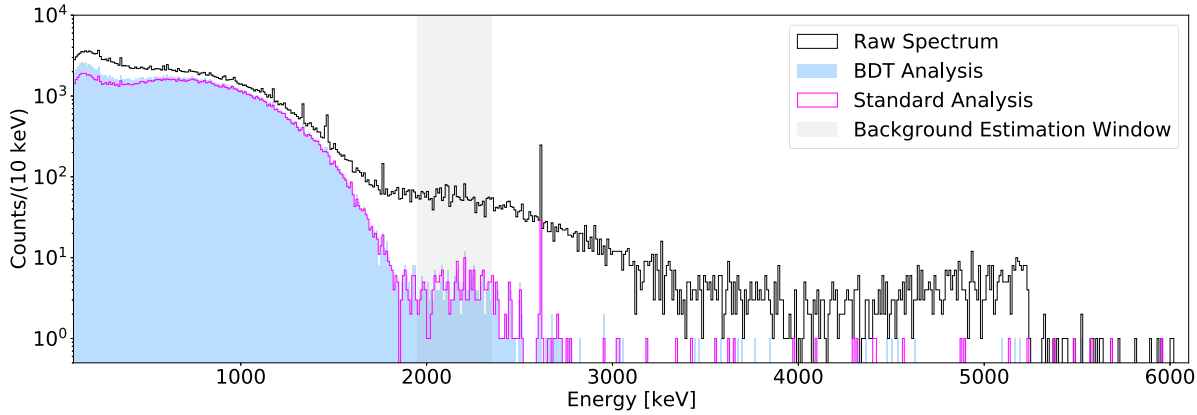


FIG. 6. Energy spectrum of the standard MAJORANA analysis and the BDT analysis with 64.5 ± 0.9 kg exposure; a good agreement is reached within the background estimation window. The low energy discrepancy between two analyses was mainly due to the `LQ` cut.

reach the same level of performance on ICPC detectors under proper tuning.

VI. MACHINE INTERPRETABILITY

We demonstrated the BDT's ability to outperform the standard MAJORANA analysis, but the source of additional classification power was not readily apparent. To identify these sources, a *post facto* machine interpretability study was performed on the trained MSBDT and αBDT. This study used a coalitional game theory concept, Shapley value [44], to interpret the decision of a BDT. The Shapley value is defined as follows:

$$\phi_i(v) = \sum_{S \subseteq N/\{i\}} \frac{|S|!(n-|S|-1)!}{n!}[v(S \cup \{i\}) - v(S)]. \quad (4)$$

v is the characteristic function that maps a subset of players to a real number. S represents a coalition of players without the player i. N is the set of all players and n is the size of that set. In this analysis, v is the BDT model mapping input features to the BDT score. Each feature is a "player" of the game. $v(S \cup \{i\}) - v(S)$ describes the difference in BDT score including/excluding feature i. This difference is summed over all coalitions S—that is, the possible combinations of all other features except $\{i\}$—to produce the Shapley value for i. Therefore, the Shapley value in the context of a BDT represents each feature's contribution to the final BDT score, assuming they work collaboratively.

The interpretability study was conducted using the SHAP package [23]. The underlying mechanism is analogous to a one-dimensional free body diagram [45]. SHAP assigns a Shapley value to each feature of the events to be interpreted. The Shapley value acts as a "force" to change the BDT score: a positive Shapley value pushes the BDT score toward a more signal-like score, while a negative Shapley value pushes the BDT score toward a more background-like score. After all "forces" are applied, the BDT reaches an equilibrium, and the equilibrium position is the BDT score of the input event. Therefore, if an event is classified as signal, the feature with the largest positive Shapley value will be the driving factor for this classification decision, while features with negative Shapley values suggest against the classification decision. An example force plot of a single MAJORANA DEMONSTRATOR event is shown in Fig. 7(c). By investigating the Shapley values on designated datasets, we will understand the driving feature which leads to the additional classification power.

A. Interpreting MSBDT

Figure 7(a) presents a summary plot to illustrate the feature importance of the MSBDT. To make this plot, we first randomly sampled 10 000 ^{228}Th DEP events and 10 000 ^{228}Th SEP events to form the interpretability dataset. The Shapley values are calculated for each event in the dataset, and the distribution of Shapley values with respect to each input feature is plotted in Fig. 7(a). The shape of these distributions represents the importance of the given feature. An important feature exhibits a dumbbell shape, indicating this feature drives the decision by a large magnitude most of the time. A less important feature exhibits a spindle shape, indicating this feature outputs a nearzero Shapley value most of the time but occasionally drives the decision with a large amplitude. An irrelevant feature exhibits a vertical bar shape, indicating that this feature almost always outputs a Shapley value of 0. Figure 7(a) ranks the importance of features from top to bottom according to this rule. The most important feature is `AvsE` as we expected, and the second most important feature is `channel`. This means MSBDT's classification power mainly comes from channel-wise calibration of `AvsE`. The least important feature is `detType` since it is redundant with `channel`. The importance ranking shown in Fig. 7(a) can also be used for feature selection. In case the computation power is limited, low-importance features such as `detType` can be removed from the input. In this work, the BDT training takes less than one minute on CPU. Therefore, low-importance features are kept since they do not seem to adversely affect the performance.

To further understand the classification power of MSBDT, especially the additional classification power compared to standard `AvsE`, we collected outperforming events from the interpretability dataset with two criteria:

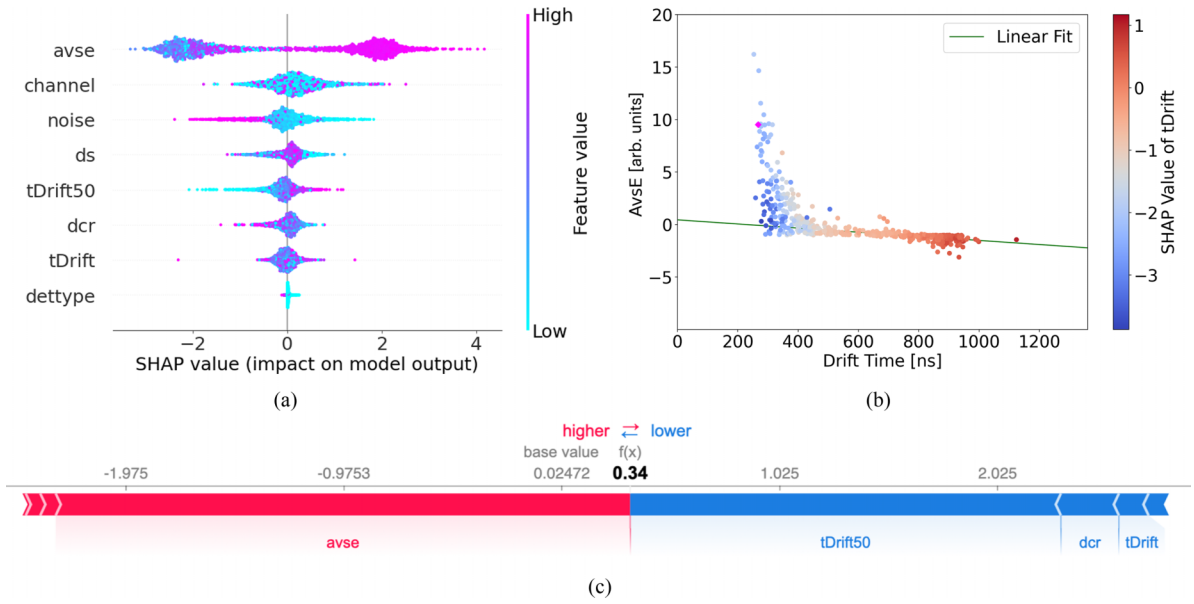


FIG. 7. (a) Feature importance plot of the MSBDT. From top to bottom, the features are ranked from the most important to the least important. The color of each dot represents the Shapley value of each feature, normalized by the highest and lowest Shapley values of all input samples. (b) The scatter plot of raw `AvsE` vs `tDrift` for outperforming events in a single detector in DS6. The color of each dot represents the sum of Shapley values assigned to both `tDrift` and `tDrift50`. Higher magnitude represents more important contributions from drift time. The green line represents a linear fit to the linear dependency of drift time. (c) Force plot of a single MAJORANA DEMONSTRATOR event denoted by the magenta diamond in (b). The Shapley values of `AvsE` and `channel` provide positive forces, while the Shapley values of `tDrift`, `tDrift50`, and `dcr` provide negative forces. The equilibrium position is at 0.34.

(i) DEP events that MSBDT classifies as signal but raw `AvsE` classifies as background,
(ii) SEP events that MSBDT classifies as background but raw `AvsE` classifies as signal.

Figure 7(b) shows the joint distribution of drift time and raw `AvsE` on a two-dimensional (2D) scatter plot on outperforming events. To avoid smearing caused by different detectors or different datasets, only outperforming events from a single detector in DS6 are shown. The color of each dot indicates the summed Shapley value of `tDrift` and `tDrift50`. Two types of drift-time dependencies are observed on raw `AvsE`: a linear dependency appears on the raw `AvsE` for large drift-time events, and a nonlinear dependency appears on low drift-time events. In the standard MAJORANA analysis, the linear dependence is corrected through a detector-by-detector drift-time correction as discussed in Sec. II. From the MSBDT's perspective, the BDT assigns a positive Shapley value on drift time to reproduce the drift-time correction: although the linear dependency leads to lower-than-usual `AvsE` at large drift time, the MSBDT successfully captures this dependency and produces a positive Shapley value to compensate for this effect. This is equivalent to the drift-time correction in standard MAJORANA analysis. Without explicit programming, the MSBDT independently learns these correlations from data and leverages them to further improve background rejection performance as expected.

The non-linear dependency happens primarily on events with drift time below 400 ns. These events happen near the point contact and drift almost immediately to it. As shown in Fig. 7(b), these fast-drifting events possess excessively high `AvsE` and will be classified as signals even if the waveform is multisite. In the standard MAJORANA analysis, we use the high `AvsE` cut and the `LQ` cut to remove these events near the point contact. In this analysis, the MSBDT assigns a negative Shapley value according to the drift time to compensate for the higher-than-usual `AvsE` values. To demonstrate this, we selected a single event from this region [the magenta diamond on Fig. 7(b)] and showed its Shapley forces in Fig. 7(c). Although the excessively high `AvsE` produces an overwhelmingly positive force, MSBDT recognizes the nonlinear drift time dependency and assigns negative forces to `tDrift` and `tDrift50` to counteract the positive force. The equilibrium position is at 0.34, which falls below the cutting threshold of MSBDT. Therefore, this event is rejected by MSBDT but accepted by standard `AvsE`. Without explicit programming, the MSBDT learns the linear and non-linear correlation from data and handles them correctly to produce better background tagging efficiency.

B. Interpreting αBDT

We used a similar approach to interpret the αBDT. Since there are only 3562 genuine alpha events, we collected all the genuine alpha events as backgrounds and 3562 randomly sampled ^{228}Th DEP events as signals to form the interpretability dataset. The summary plot is shown in Fig. 8(a). As expected, raw `DCR` is the most important feature in making a classification decision. `tDrift50` is the second most important feature, indicating that the αBDT is mainly performing a drift-time correction on raw `DCR` to enhance its classification power.

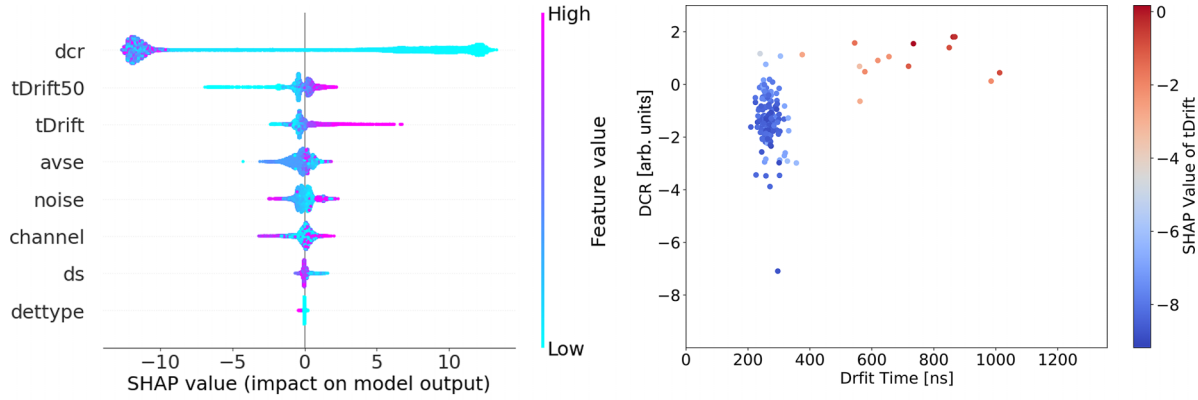


FIG. 8. (a) Feature importance plot of the αBDT. (b) The scatter plot of raw DCR vs tDrift for all outperforming genuine alpha events. The color of each dot represents the sum of Shapley values assigned to both tDrift and tDrift50. Higher magnitude represents more important contributions from drift time.

Similarly to the MSBDT, detType is the least important feature since it is redundant with channel.

Outperforming events are collected from the interpretability dataset to understand the additional classification power of the αBDT. Since the signal sacrifice of the αBDT and DCR is negligible, the outperforming dataset is defined as genuine alpha events rejected by the αBDT but accepted by the standard DCR. Figure 8(b) shows the joint distribution of drift time and standard DCR on a 2D scatter plot on outperforming events. These events form a cluster near a drift time of 200 ns, indicating that they are surface alpha events near the point contact. After creation, these events drift to the point contact almost immediately, leaving almost no delayed charge on the passivated surface, thus violating the first principle of the DCR cut. However, the fast-drifting nature allows the αBDT to efficiently tag these events based on their drift time, thus outperforming the traditional analysis. As αBDT interpretability study revealed the importance of these backgrounds, a dedicated high AvsE cut is introduced into the standard MAJORANA analysis to reject them. High AvsE turns out to also reject multisite event near the point contact as we discussed in Sec. VI A. The interpretability study also shows that tDrift50 is more important in the αBDT model than tDrift. This can be explained by the difference between the calculation of these two parameters. A typical outperforming event is shown in Fig. 9. When a surface alpha event happens near the point contact, the charge deposition starts almost immediately, leading to a sharp rising edge of the waveform. On the other hand, the passivated surface reduces the drift speed of charges comparatively further away from the point contact. This effect delays the completion of charge deposition, leading to a rounded top of the waveform. Since tDrift is calculated from the the start of the rise to the time when the waveform reaches 99% of its maximum amplitude, the rounded-top structure significantly increases the value of tDrift, allowing it to appear as a slowly drifting event, thus escaping the low-drift-time/high-AvsE cut. However, tDrift50 is immune to the rounded-top structure since it is calculated only up to 50% of the waveform amplitude. The interpretability study suggests that a tDrift50-based cut could be developed to further benefit the alpha rejection in future first-principle analyses.

The interpretability study allows machine learning analysis to unveil physics in germanium detectors. Leveraging multivariate correlations and automatic categorization, the BDT was able to outperform individual PSD parameters and match both the GAT and the ORNL analyses, as discussed in Sec. V with less detector-by-detector calibration. Furthermore, the interpretability study leverages the additional classification power to reveal the importance of new background categories. This eventually led to the implementation of a high AvsE cut in the standard MAJORANA analysis and suggests a new direction for future improvement. The reciprocal relationship between the machine learning analysis and the traditional, first-principle analysis revealed by the interpretability study demonstrates that an interpretable machine learning analysis can not only outperform but also benefit the traditional analysis.

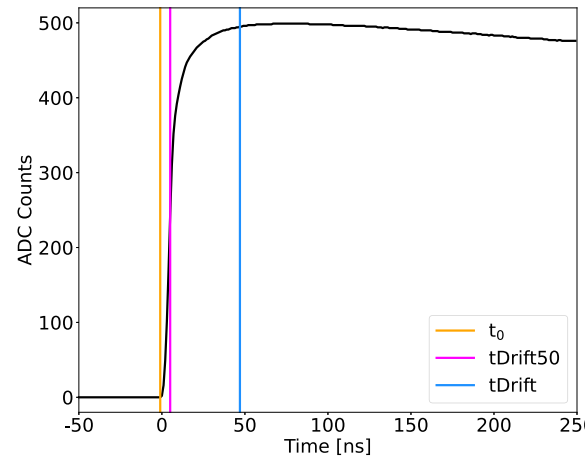


FIG. 9. A typical surface alpha event waveform (black) near the point contact of the MAJORANA PPC detector. t_0 is the start of charge deposition. The time intervals of tDrift50 and tDrift features are shown.

VII. CONCLUSION

In this work, we have presented the first machine learning analysis for the MAJORANA DEMONSTRATOR; this is also the first interpretable machine learning analysis of any germanium detector experiment. This analysis contains two parts: learning from the data to improve background rejections and learning from the machine to understand classification power. Leveraging gradient boosted decision trees and data augmentation, this analysis outperforms the the individual PSD parameters and matches the overall results of the highly optimized standard MAJORANA analysis [32]. Learning from data also closes the gap between two independently developed analyses applied to different types of detectors.

For the first time in the field, a thorough machine interpretability study is conducted, leveraging the Shapley value in coalitional game theory. This study not only justifies BDT's capability to capture multivariate correlations but also to independently discover new background categories to reveal its importance. The machine learning analysis and the standard MAJORANA analysis established a reciprocal relationship through the interpretability study. Since BDT model is widely used in the particle and nuclear physics community [46–52], this work provides a template for interpreting the BDT model to gain more physical insight and even make new scientific discoveries.

This work has focused on developing and interpreting the first machine learning analysis for the MAJORANA DEMONSTRATOR. The data-driven nature of this analysis allows a straightforward generalization to different germanium detector experiments, especially the next-generation tonne-scale experiment LEGEND-1000 [22]. Given the large number of detectors, detector- and run-level tuning may be time consuming in LEGEND. In that case, the BDT's ability to simultaneously train on all detectors would be highly beneficial. Furthermore, the interpretability study allows us to unravel the black-box nature of machine learning models to reveal underlying physics and independently discover new background categories without explicit programming. We intend to apply this model to LEGEND data, which could enable improvements in background rejection, and possibly help us gain a more nuanced understanding of the detector performance. Our future work involves using more powerful and versatile machine learning models such as recurrent neural network (RNN). RNN can be trained directly on the DEMONSTRATOR's waveform, which opens up an entirely new avenue for more machine learning applications.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics under Contracts/Awards No. DE-AC02-05CH11231, No. DE-AC05-00OR22725, No. DE-AC05-76RL0130, No. DE-FG02-97ER41020, No. DE-FG02-97ER41033, No. DE-FG02-97ER41041, No. DE-SC0012612, No. DE-SC0014445, No. DE-SC0018060, No. DE-SC0022339, and LANLEM77/LANLEM78. We acknowledge support from the Particle Astrophysics Program and Nuclear Physics Program of the National Science Foundation through Grants No. MRI-0923142, No. PHY-1003399, No. PHY-1102292, No. PHY-1206314, No. PHY-1614611, No. PHY-1812409, No. PHY-1812356, No. PHY-2111140, and No. PHY-2209530. We gratefully acknowledge the support of the Laboratory Directed Research & Development (LDRD) program at Lawrence Berkeley National Laboratory for this work. We gratefully acknowledge the support of the U.S. Department of Energy through the Los Alamos National Laboratory LDRD Program and through the Pacific Northwest National Laboratory LDRD Program for this work. We gratefully acknowledge the support of the South Dakota Board of Regents Competitive Research Grant. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, funding Reference No. SAPIN-2017-00023, and from the Canada Foundation for Innovation John R. Evans Leaders Fund. This research used resources provided by the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory and by the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility at Lawrence Berkeley National Laboratory. We thank our hosts and colleagues at the Sanford Underground Research Facility for their support.

[1] M. Agostini, G. Benato, J. A. Detwiler, J. Menéndez, and F. Vissani, Toward the discovery of matter creation with neutrinoless double-beta decay, arXiv:2202.01787

[2] M. J. Dolinski, A. W. P. Poon, and W. Rodejohann, Neutrinoless double-beta decay: Status and prospects, Annu. Rev. Nucl. Part. Sci. **69**, 219 (2019).

[3] S. Dell'Oro, S. Marcocci, M. Viel, and F. Vissani, Neutrinoless double beta decay: 2015 review, Adv. High Energy Phys. **2016**, 2162659 (2016).

[4] M. Fukugita and T. Yanagida, Baryogenesis without grand unification, Phys. Lett. B **174**, 45 (1986).

[5] S. Abe, S. Asami, M. Eizuka, S. Futagi *et al.* (KamLAND-Zen Collaboration), First search for the majorana nature of neutrinos in the inverted mass ordering region with KamLAND-Zen, arXiv:2203.02139.

[6] M. Agostini, G. R. Araujo, A. M. Bakalyarov, M. Balata, I. Barabanov, L. Baudis, C. Bauer *et al.* (GERDA Collaboration), Final Results of GERDA on the Search for Neutrinoless Double-β Decay, Phys. Rev. Lett. **125**, 252502 (2020).

[7] A. Li, Z. Fu, L. Winslow, C. Grant, H. Song, H. Ozaki, I. Shimizu, and A. Takeuchi, KamNet: An integrated spatiotemporal deep neural network for rare event search in KamLAND-Zen, arXiv:2203.01870.

[8] A. Li, A. Elagin, S. Fraker, C. Grant, and L. Winslow, Suppression of cosmic muon spallation backgrounds in liquid scintillator detectors using convolutional neural networks, Nucl. Instrum. Methods Phys. Res., Sect. A **947**, 162604 (2019).

[9] S. Hayashida, Application of the RNN in the fundamental physics with KamLAND experiment, Am. J. Comput. Sci. Inf. Technol. (2019), doi: 10.21767/2349-3917-C1-009.

[10] J. Renner, A. Farbin, J. M. Vidal, J. Benlloch-Rodríguez, A. Botas, P. Ferrario, J. Gómez-Cadenas *et al.* (NEXT Collaboration), *Background Rejection in NEXT Using Deep Neural Networks* (IOP, Bristol, 2017), pp. T01004–T01004.

[11] R. Acciarri, C. Adams, R. An, J. Asaadi, M. Auger, L. Bagby *et al.* (MicroBooNE Collaboration), *Convolutional Neural Networks Applied to Neutrino Events in a Liquid Argon Time Projection Chamber* (IOP, Bristol, 2017), pp. P03011–P03011.

[12] A. Aurisano, A. Radovic, D. Rocco, A. Himmel, M. Messier, E. Niner, G. Pawloski, F. Psihas, A. Sousa, and P. Vahle, *A Convolutional Neural Network Neutrino Event Classifier* (IOP, Bristol, 2016), pp. P09001–P09001.

[13] P. Holl, L. Hauertmann, B. Majorovits, O. Schulz, M. Schuster, and A. J. Zsigmond, Deep learning based pulse shape discrimination for germanium detectors, Eur. Phys. J. C **79**, 450 (2019).

[14] F. Psihas, M. Groh, C. Tunnell, and K. Warburton, A review on machine learning for neutrino experiments, Int. J. Mod. Phys. A **35**, 2043005 (2020).

[15] A. Ashtari Esfahani *et al.*, Cyclotron radiation emission spectroscopy signal classification with machine learning in project 8, New J. Phys. **22**, 033004 (2020).

[16] K. Liu, J. Greitemann, and L. Pollet, Learning multiple order parameters with interpretable machines, Phys. Rev. B **99**, 104410 (2019).

[17] P. Ponte and R. G. Melko, Kernel methods for interpretable machine learning of order parameters, Phys. Rev. B **96**, 205146 (2017).

[18] P. Y. Lu, S. Kim, and M. Soljačić, Extracting Interpretable Physical Parameters from Spatiotemporal Systems Using Unsupervised Learning, Phys. Rev. X **10**, 031056 (2020).

[19] C. E. Aalseth *et al.* (MAJORANA Collaboration), Search for Neutrinoless Double-β Decay in ^{76}Ge with the MAJORANA DEMONSTRATOR, Phys. Rev. Lett. **120**, 132502 (2018).

[20] N. Abgrall *et al.* (MAJORANA Collaboration), The MAJORANA DEMONSTRATOR neutrinoless double-beta decay experiment, Adv. High Energy Phys. **2014**, 365432 (2014).

[21] S. I. Alvis *et al.* (MAJORANA Collaboration), A Search for Neutrinoless double-beta decay in ^{76}Ge with 26 kg yr of exposure from the MAJORANA DEMONSTRATOR, Phys. Rev. C **100**, 025501 (2019).

[22] N. Abgrall *et al.* (LEGEND Collaboration), The Large Enriched Germanium Experiment for Neutrinoless $\beta\beta$ Decay: LEGEND-1000 Preconceptual Design Report, arXiv:2107.11462.

[23] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. **2**, 56 (2020).

[24] F. Wang, S. Huang, R. Gao, Y. Zhou, C. Lai, Z. Li, W. Xian, X. Qian, Z. Li, Y. Huang *et al.*, Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility, Cell Discovery **6**, 1 (2020).

[25] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang *et al.*, An interpretable mortality prediction model for COVID-19 patients, Nat. Mach. Intell. **2**, 283 (2020).

[26] Y. Chen, L. Ouyang, F. S. Bao, Q. Li, L. Han, B. Zhu, M. Xu, J. Liu, Y. Ge, and S. Chen, An interpretable machine learning framework for accurate severe vs non-severe COVID-19 clinical type classification, doi:10.1101/2020.05.18.20105841.

[27] J. Jiménez-Luna, F. Grisoni, and G. Schneider, Drug discovery with explainable artificial intelligence, Nat. Mach. Intell. **2**, 573 (2020).

[28] S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit, and H. J. Kulik, Understanding the diversity of the metal-organic framework ecosystem, Nat. Commun. **11**, 4068 (2020).

[29] R. S. Baker and A. Hawn, Algorithmic bias in education, EdArXiv:10.35542/osf.io/pbmvz.

[30] N. Abgrall *et al.* (MAJORANA Collaboration), The processing of enriched germanium for the MAJORANA DEMONSTRATOR and R&D for a next generation double-beta decay experiment, Nucl. Instrum. Methods Phys. Res., Sect. A **877**, 314 (2018).

[31] R. Cooper, D. Radford, P. Hausladen, and K. Lagergren, A novel HPGe detector for gamma-ray tracking and imaging, Nucl. Instrum. Methods Phys. Res., Sect. A **665**, 25 (2011).

[32] I. J. Arnquist *et al.* (MAJORANA Collaboration), Final result of the MAJORANA DEMONSTRATOR's search for neutrinoless double-β decay in ^{76}Ge, arXiv:2207.07638.

[33] N. Abgrall *et al.* (Majorana Collaboration), The MAJORANA DEMONSTRATOR readout electronics system, J. Instrum. **17**, T05003 (2022).

[34] N. Abgrall *et al.* (MAJORANA Collaboration), ADC Nonlinearity Correction for the MAJORANA DEMONSTRATOR, IEEE Trans. Nucl. Sci. **68**, 359 (2021).

[35] N. Abgrall *et al.* (MAJORANA Collaboration), The Majorana Demonstrator calibration system, Nucl. Instrum. Methods Phys. Res., Sect. A **872**, 16 (2017).

[36] I. J. Arnquist *et al.* (MAJORANA Collaboration), Charge Trapping and energy performance of the MAJORANA DEMONSTRATOR, arXiv:2208.03424.

[37] S. I. Alvis *et al.* (MAJORANA Collaboration), Multisite event discrimination for the MAJORANA DEMONSTRATOR, Phys. Rev. C **99**, 065501 (2019).

[38] I. J. Arnquist *et al.* (MAJORANA Collaboration), α-event characterization and rejection in point-contact HPGe detectors, Eur. Phys. J. C **82**, 226 (2022).

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. **16**, 321 (2002).

[40] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in *Advances in Neural Information Processing Systems*, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Red Hook, NY, 2017).

[41] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. **29**, 1189 (2001).

[42] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization, in *Advances in Neural Information Processing Systems*, Vol. 33, edited by H. Larochelle and M. Ranzato and R. Hadsell and M.F. Balcan and H. Lin (Curran Associates, Red Hook, NY, 2020).

[43] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition **30**, 1145 (1997).

[44] L. S. Shapley, in *Quota Solutions of N-Person Games*, edited by E. Artin and M. Morse, 343 (Rand, Santa Monica, 1953), p. 343.

[45] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim *et al.*, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biom. Eng. **2**, 749 (2018).

[46] P. Abratenko *et al.* (MicroBooNE Collaboration), Search for heavy neutral leptons decaying into muon-pion pairs in the MicroBooNE detector, Phys. Rev. D **101**, 052001 (2020).

[47] CMS Collaboration, Observation of same-sign WW production from double parton scattering in proton-proton collisions at $\sqrt{s} = 13$ TeV, arXiv:2206.02681.

[48] G. Aad *et al.* (ATLAS Collaboration), Measurements of Higgs boson production and couplings in the four-lepton channel in *pp* collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector, Phys. Rev. D **91**, 012006 (2015).

[49] M. Aaboud *et al.* (ATLAS Collaboration), Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s} = 13$ TeV arXiv:1805.01845, p. 089.

[50] A. M. Sirunyan *et al.* (CMS Collaboration), Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV J. Instrumentation **15**, P10017 (2020).

[51] A. A. Aguilar-Arevalo *et al.* (MiniBooNE Collaboration), A Search for Electron Neutrino Appearance at the $\Delta m^2 \sim 1$ eV^2 Scale, Phys. Rev. Lett. **98**, 231801 (2007).

[52] V. M. Abazov *et al.* (D0 Collaboration), Evidence for production of single top quarks, Phys. Rev. D **78**, 012005 (2008).