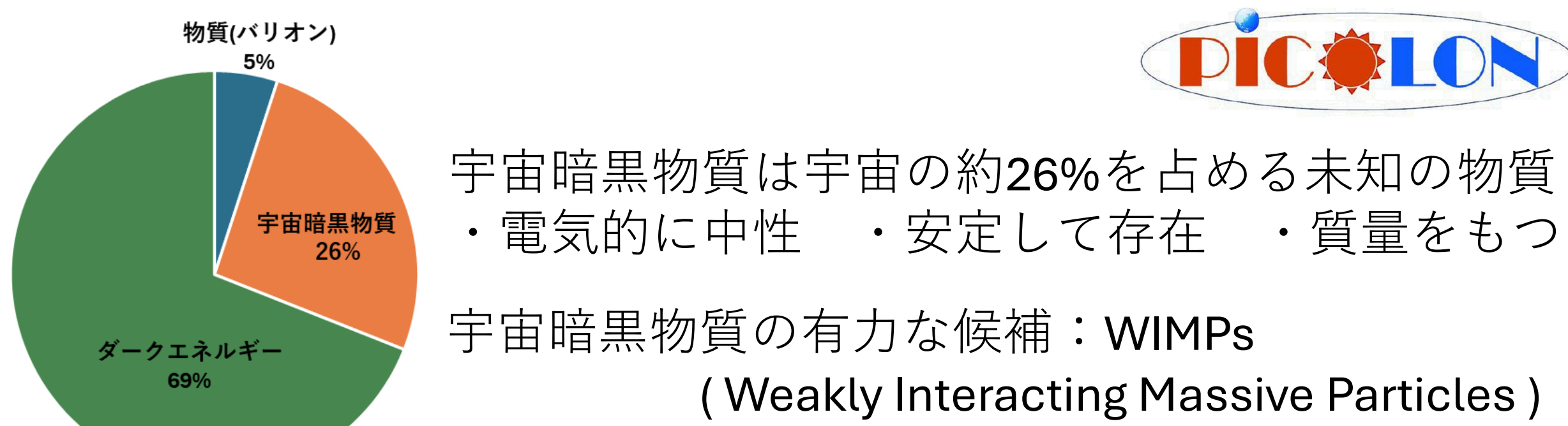


I. PICOLON プロジェクトによる宇宙暗黒物質の探索

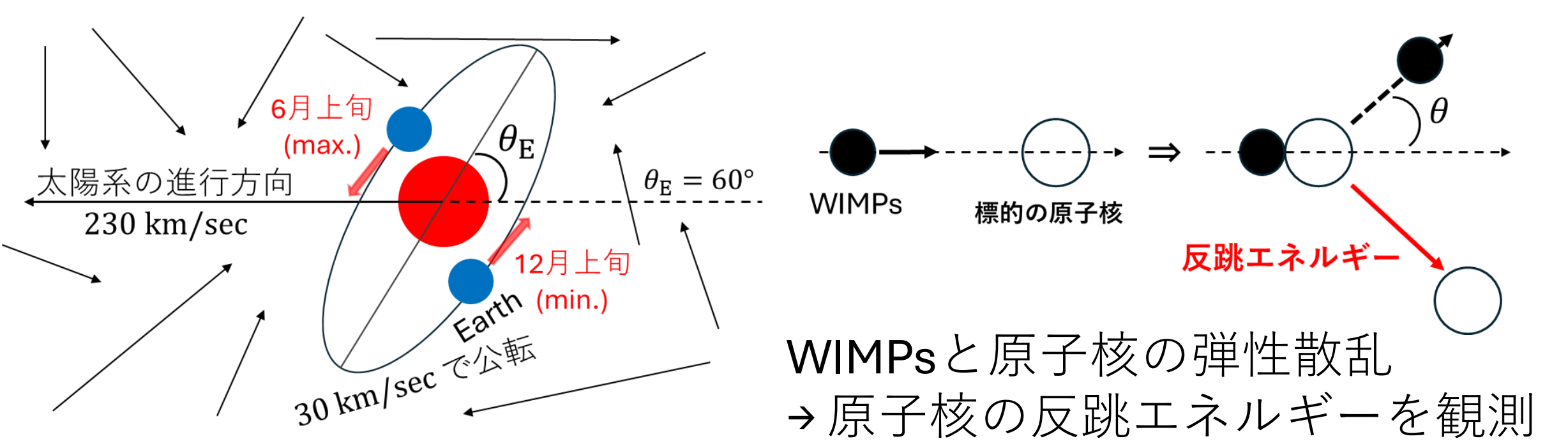
Pure Inorganic Crystal Observatory for LOW-energy Neutr(al)ino



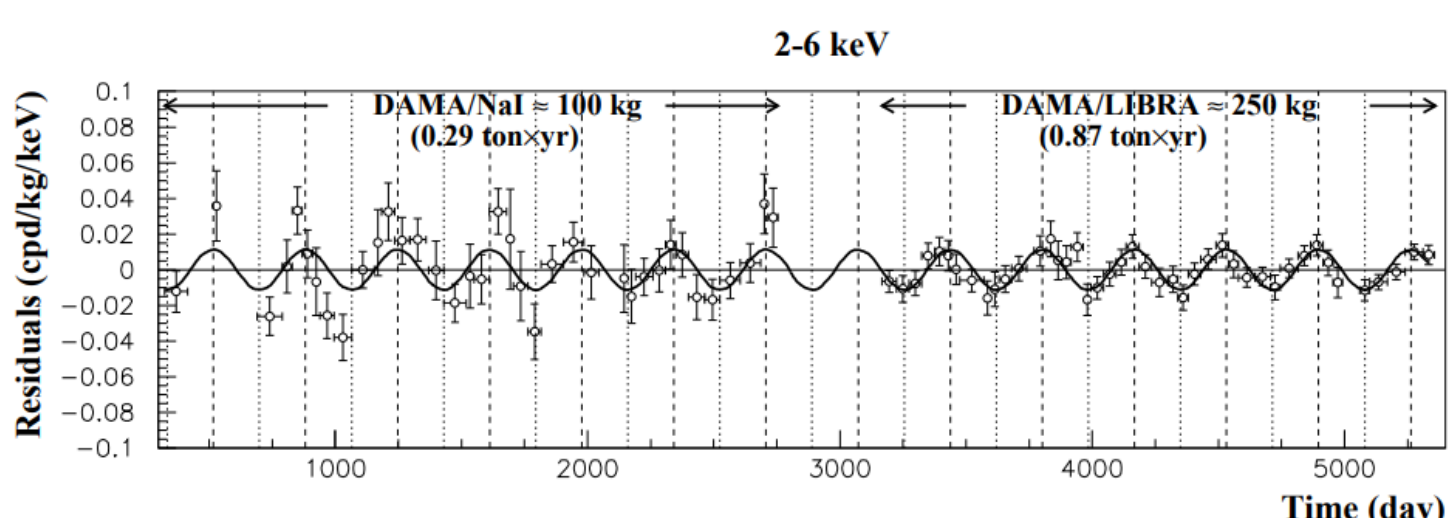
宇宙暗黒物質は宇宙の約26%を占める未知の物質
・電的に中性 ・安定して存在 ・質量をもつ

宇宙暗黒物質の有力な候補: WIMPs
(Weakly Interacting Massive Particles)

目標: 超高純度 NaI(Tl) 検出器を開発し, WIMPsの相互作用を総合的に観測する



DAMA/LIBRAが報告した季節変動



→ この結果を検証を行う。

↓
WIMPsの事象は非常に稀
→ 実験感度の向上が必要!

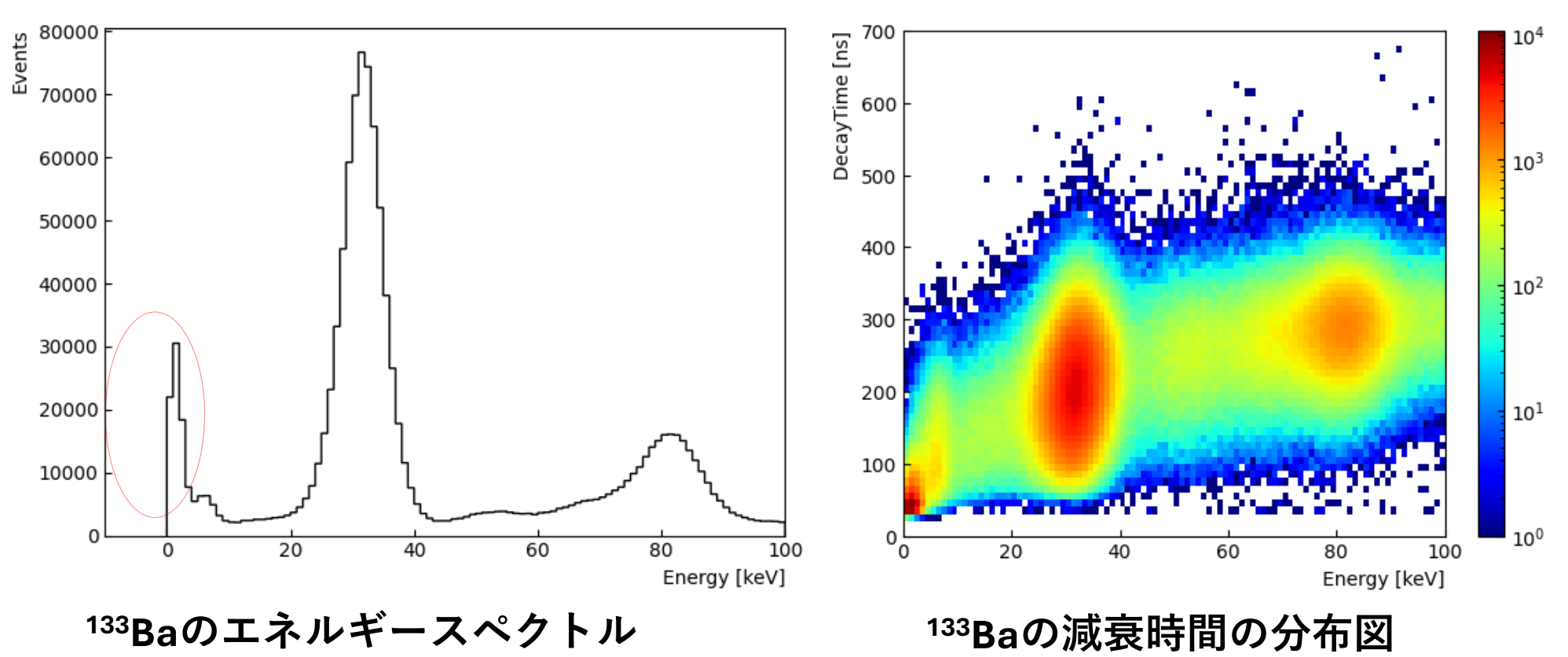
Proceedings to the 16th Bled Workshop "What Comes Beyond the Standard Models?" (Dec. 2013) Vol.14 No.2
より引用

[1]より

現在の課題: 実験感度の向上

・バックグラウンド濃度の低減

低エネルギー側にノイズが多く存在

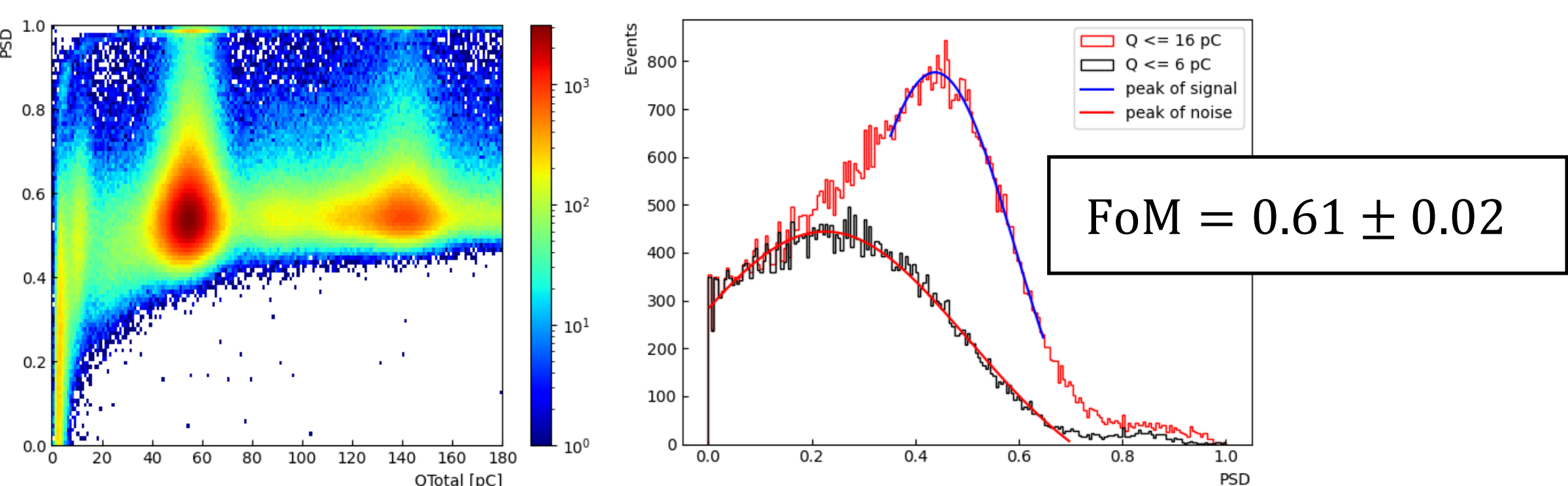


従来のノイズ除去法

・NaIの時定数が約200 nsであることを利用

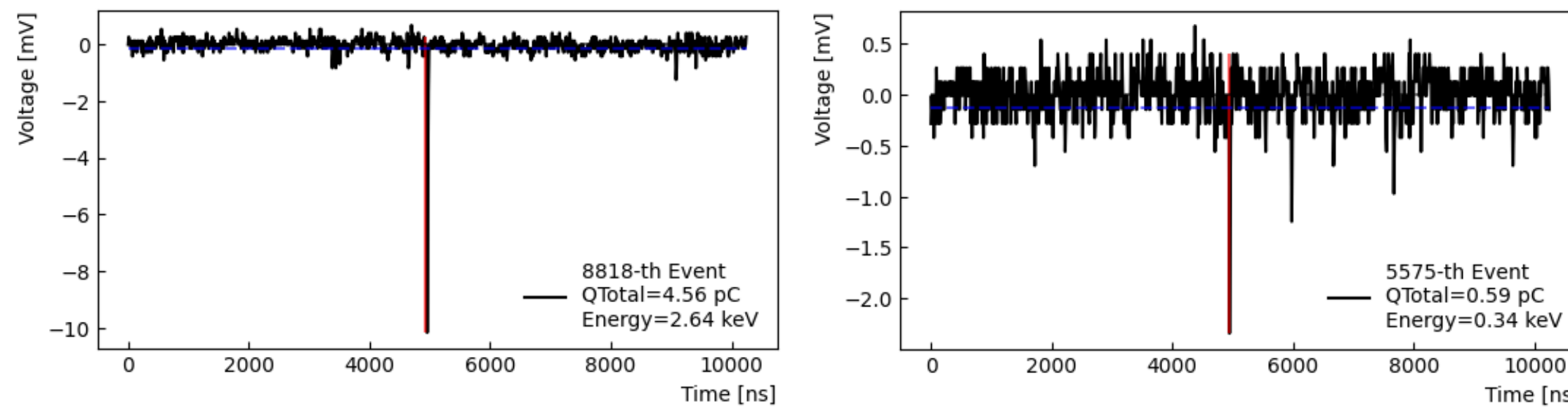
$$PSD = \frac{Q_{Part}}{Q_{Total}} = \frac{\int_{200\text{ ns}}^{1200\text{ ns}} Idt}{\int_{0\text{ ns}}^{1200\text{ ns}} Idt}$$

$$FoM = \frac{|\mu_N - \mu_S|}{\sqrt{\sigma_N^2 + \sigma_S^2}}$$

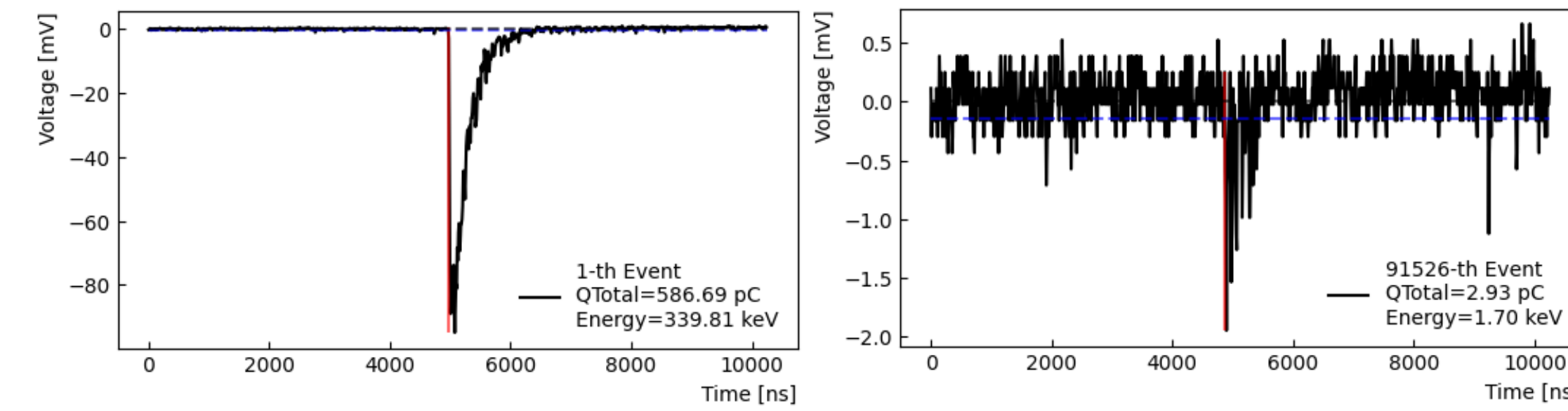


※電荷 Q = 16 pC ~ Energy = 10 keV

・ノイズの波形

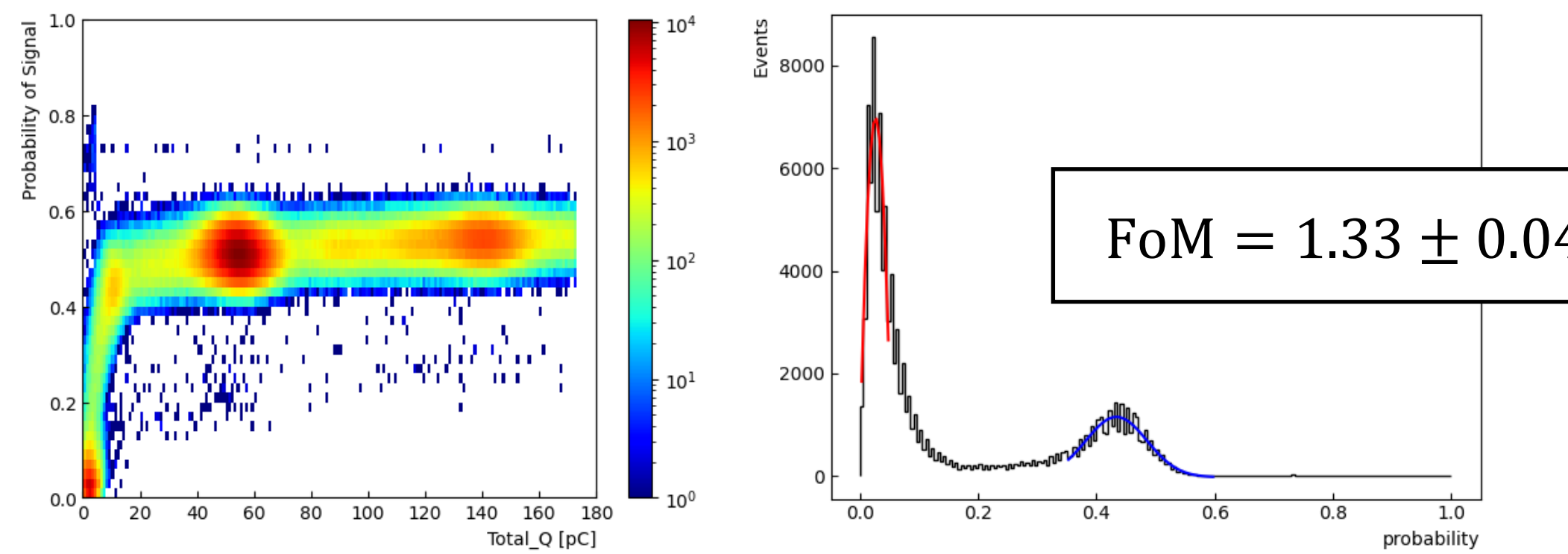


・シグナルの波形



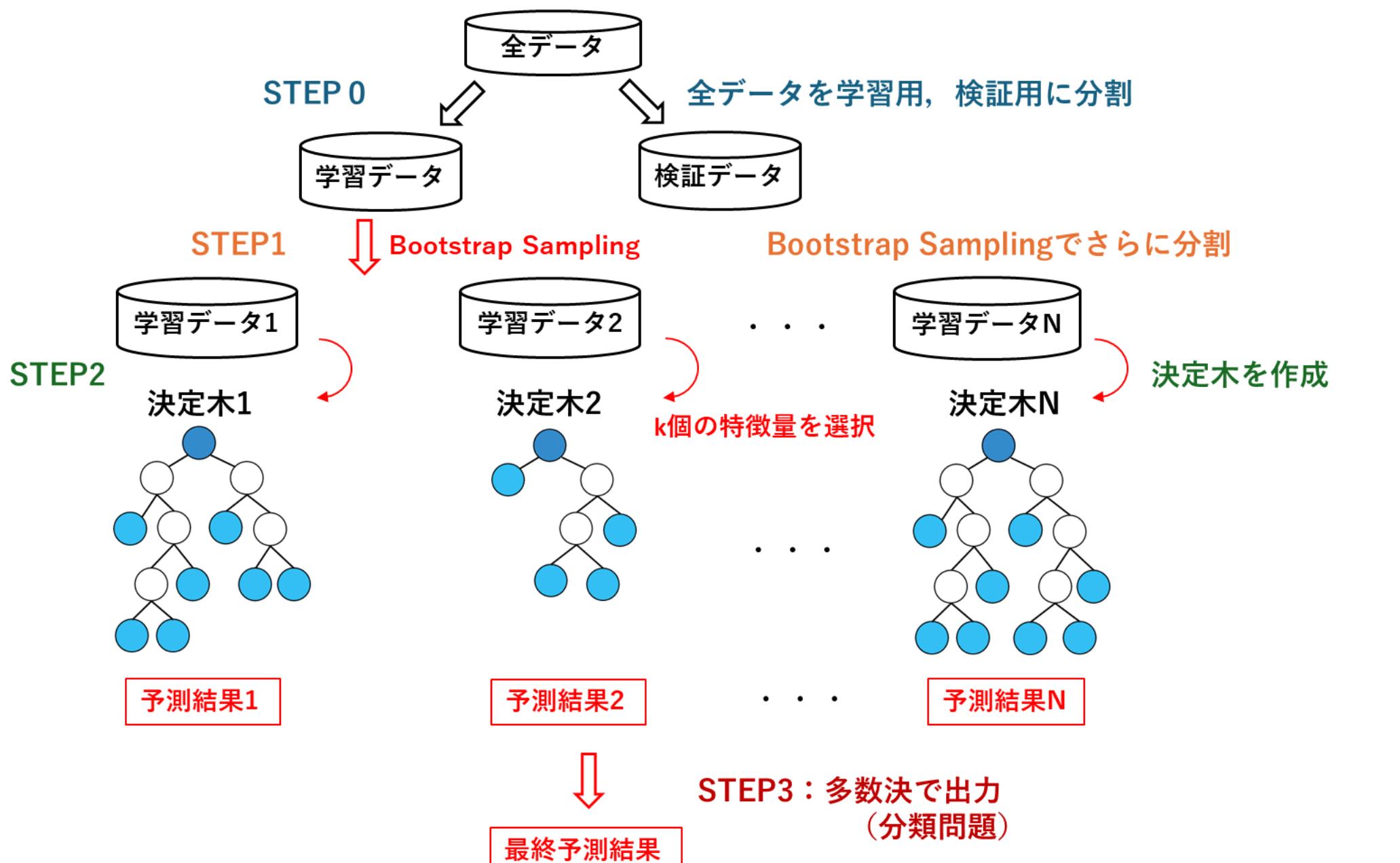
→ 波形のデータ (1024個) を機械学習し, ノイズを除去する

機械学習の有用性 (ランダムフォレストを使用)



II. 使用する機械学習アルゴリズムとデータセット

ランダムフォレスト



▶ Bootstrap Sampling

全データからランダムに一部のデータを取り出して学習に用いる方法

▶ 確率の出力方法

各決定木のリーフノードに到達したサンプルのクラス分布を基にクラス確率を計算し, それらの確率を平均

$$P(c | j) = \frac{N_c}{N_j}$$

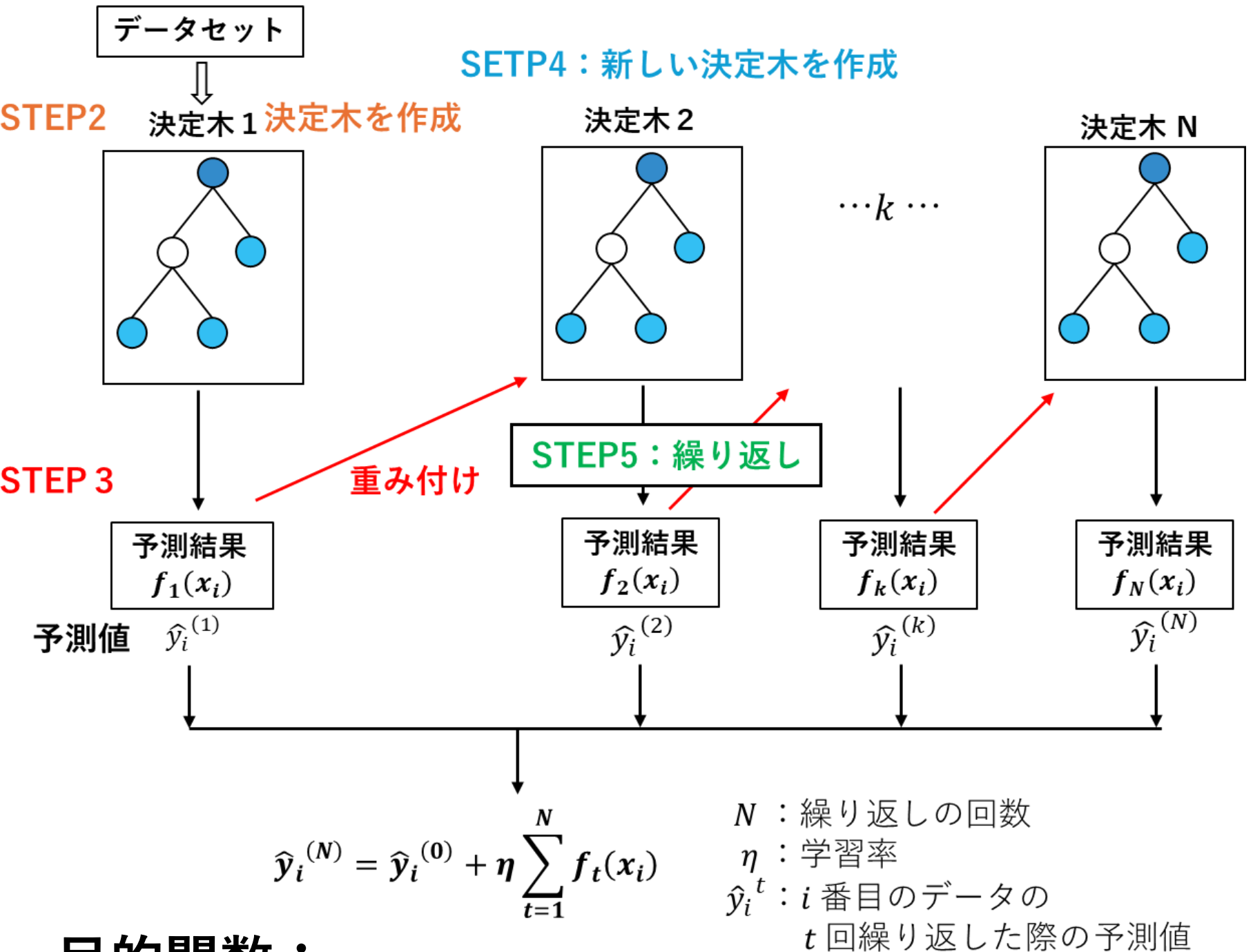
c: クラス, j: リーフノードの番号, N: サンプル数

▶ 最大の特徴: 決定木が独立に, 並列に作成 & 学習

[2]より

XGBoost (勾配ブースティング決定木)

STEP1: 初期値を設定 ($\hat{y}^{(0)} = 0.5$)



目的関数:

$$L = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

損失関数: $l(y_i, \hat{y}_i) = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$
正則化項: $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 + \alpha \sum_{j=1}^T |\omega_j|$

▶ 確率の出力方法

予測値をシグモイド関数で確率に変換: $\hat{p}_i = \sigma(\hat{y}_i)$

▶ 最大の特徴

決定木が直列に作成 & 学習

勾配:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i} = \hat{p}_i - y_i$$

ヘッシアン:

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} = \hat{p}_i(1 - \hat{p}_i)$$

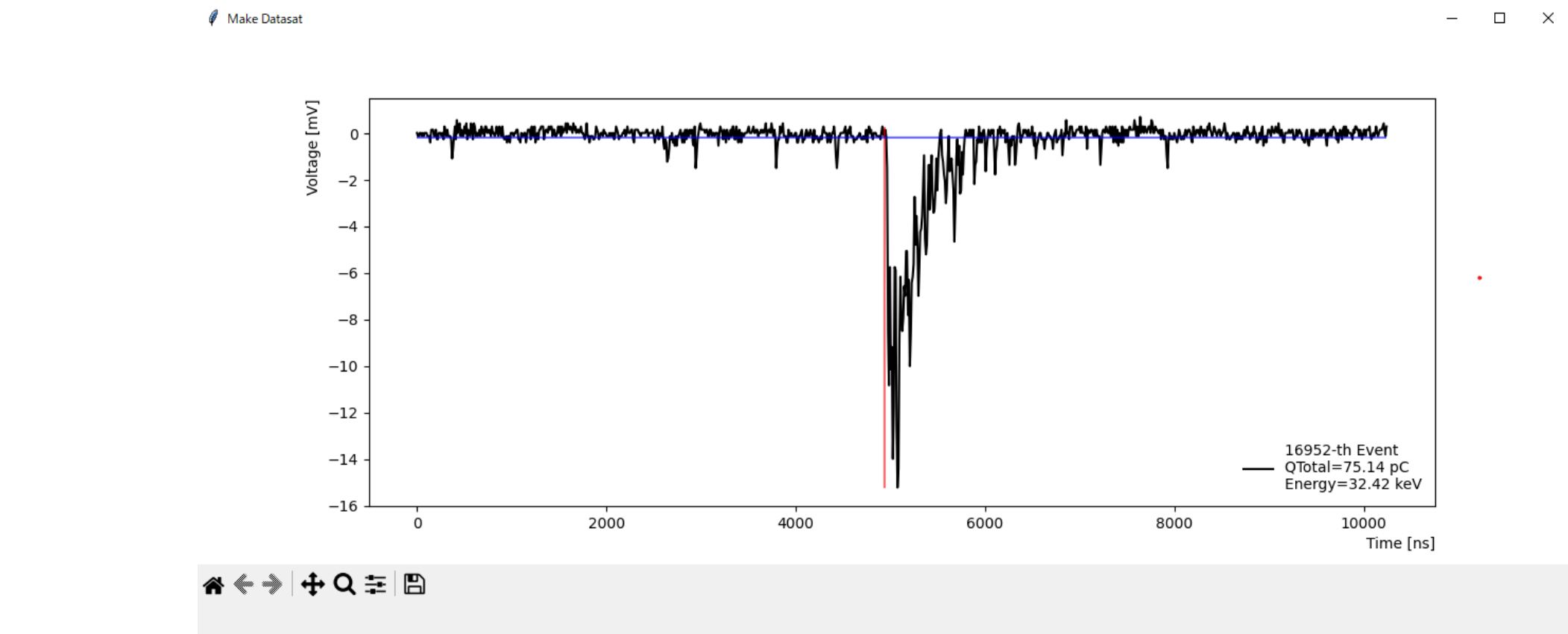
y_i : 実際のラベル
 $\hat{y}_i$: 予測値
 λ : L2正則化の制御
 α : L1正則化の制御
 T : リーフノードの数
 ω_j : リーフノードの重み

[3]より

データセット

▶ 作成方法

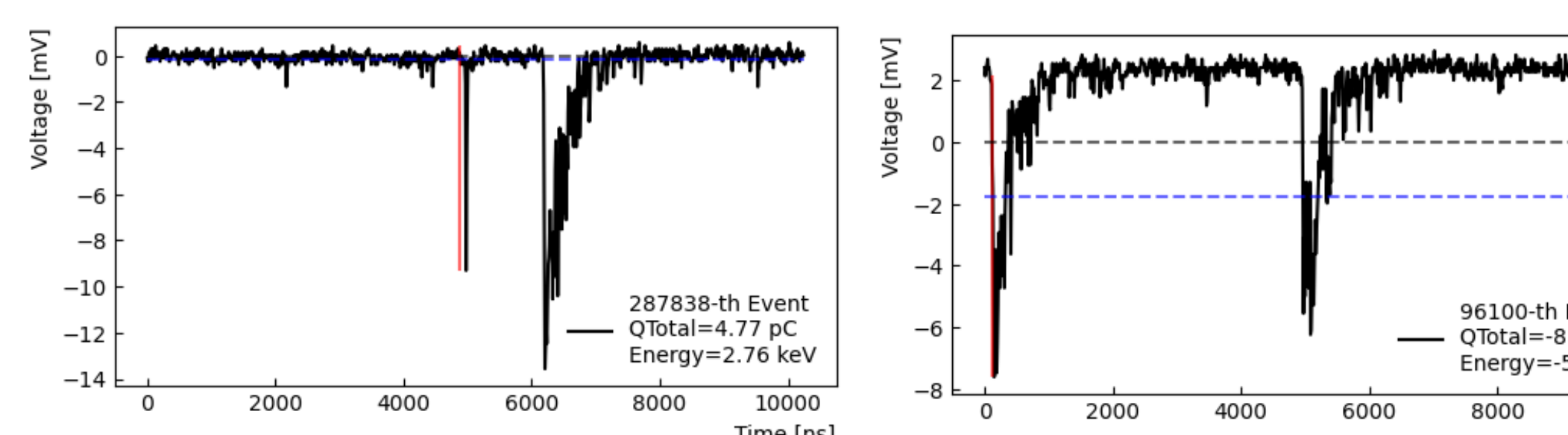
GUIを作成して, 1つ1つ見てシグナルなら0, ノイズなら1を入力する



▶ 作成したデータセット

データの個数: 10910個 内訳: {0: 7002, 1: 3908}

▶ データセットに入れなかった波形

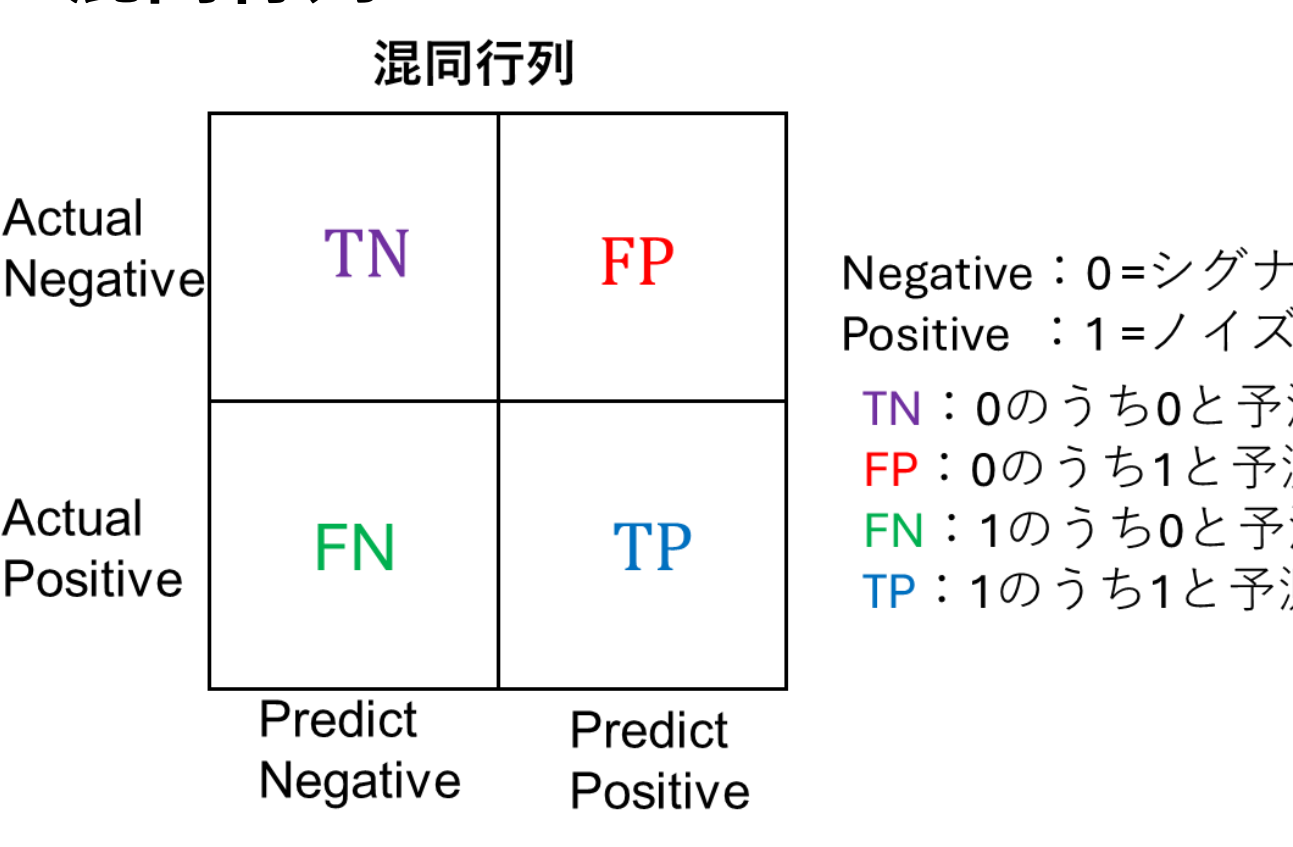


ダブルパルスやベースラインが浮いているもの

III. ランダムフォレスト vs XGBoost

モデルの評価方法

▶ 混同行列



$$FPR = \frac{FP}{FP + TN}, \quad TPR = \frac{TP}{TP + FN}$$

▶ ROC AUC

予測確率に対して閾値を変更してFPRとTPRをそれぞれ計算

横軸 FPR, 縦軸 TPR のグラフを書く

左上 (FPR, TPR) = (0, 1) の近くを通るほどモデルの精度が高い

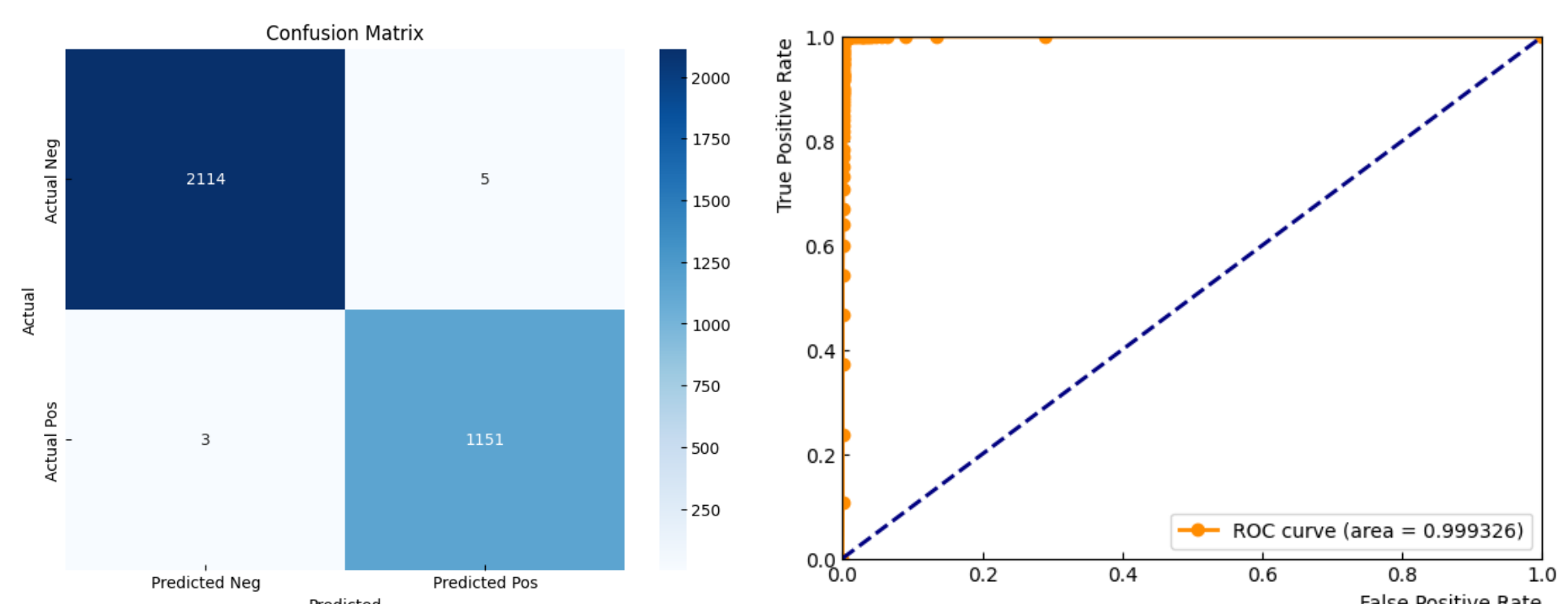
▶ Log Loss

$$l = -(y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i))$$

$$\hat{p}_i = \sigma(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}}$$

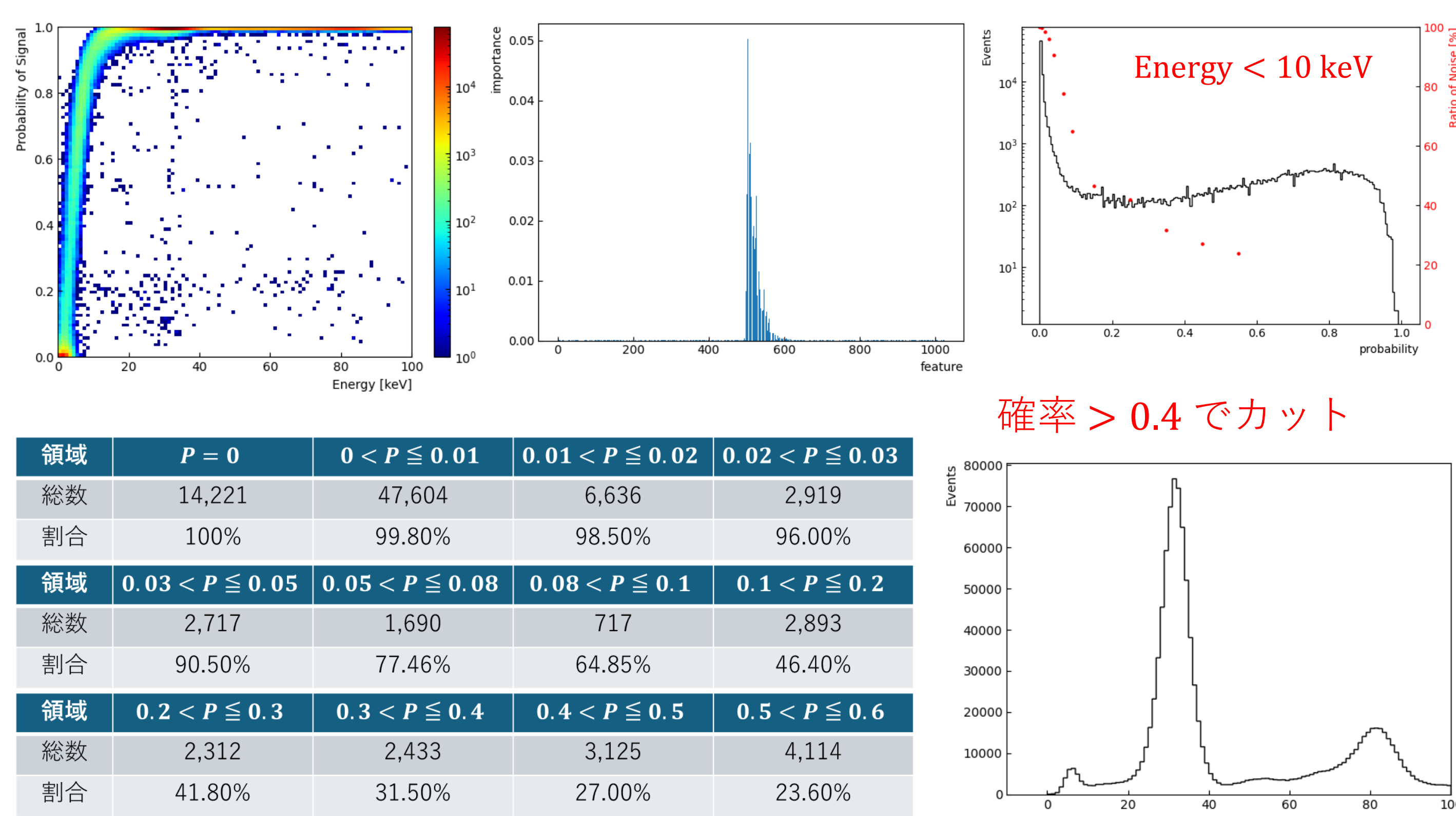
ランダムフォレスト

▶ モデルの評価



Accuracy: 0.9976
ROC AUC: 0.99933

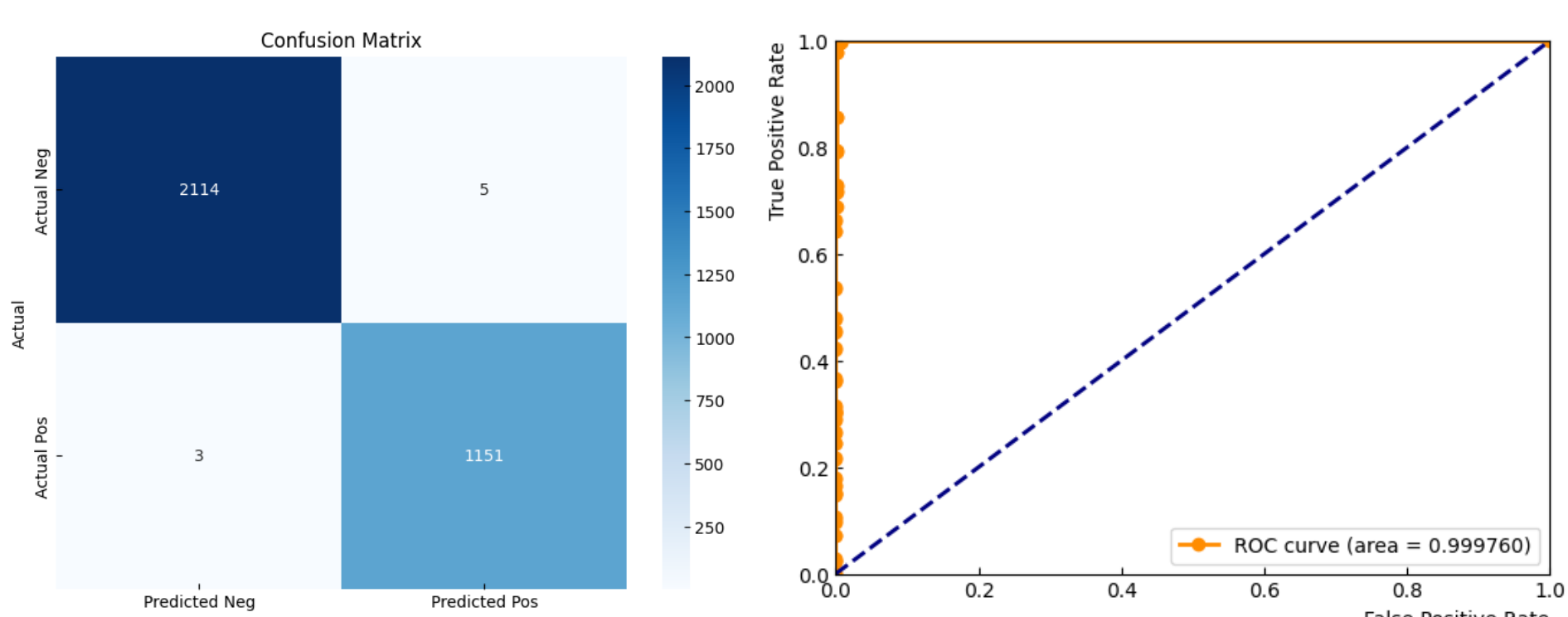
▶ 機械学習の結果



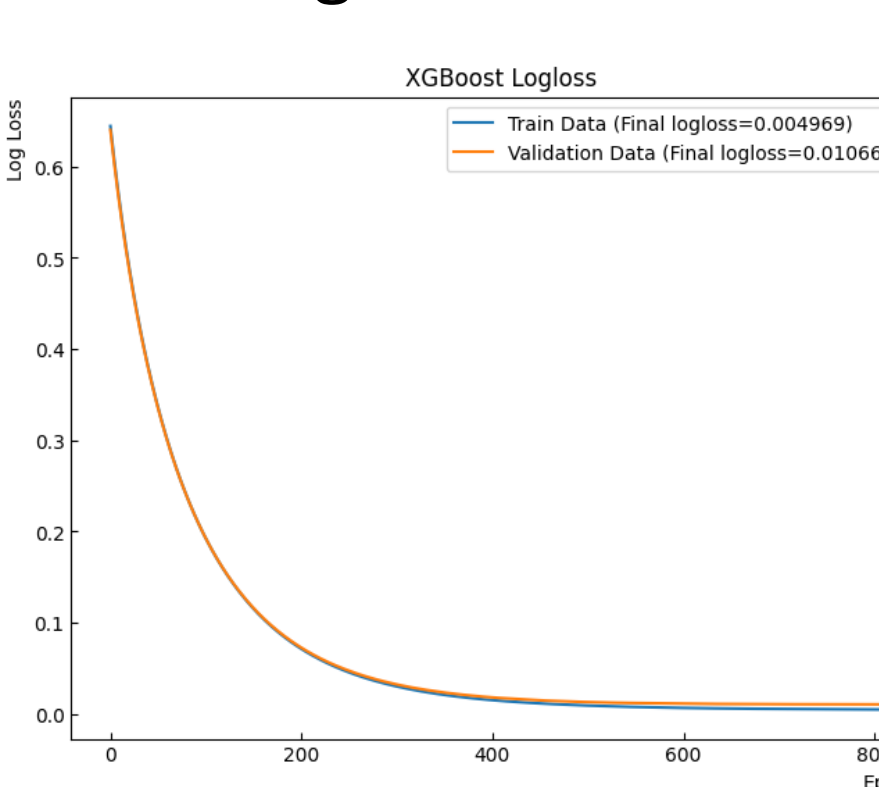
領域	$P = 0$	$0 < P \leq 0.01$	$0.01 < P \leq 0.02$	$0.02 < P \leq 0.03$
総数	14,221	47,604	6,636	2,919
割合	100%	99.80%	98.50%	96.00%
領域	$0.03 < P \leq 0.05$	$0.05 < P \leq 0.08$	$0.08 < P \leq 0.1$	$0.1 < P \leq 0.2$
総数	2,717	1,690	717	2,893
割合	90.50%	77.46%	64.85%	46.40%
領域	$0.2 < P \leq 0.3$	$0.3 < P \leq 0.4$	$0.4 < P \leq 0.5$	$0.5 < P \leq 0.6$
総数	2,312	2,433	3,125	4,114
割合	41.80%	31.50%	27.00%	23.60%

XGBoost

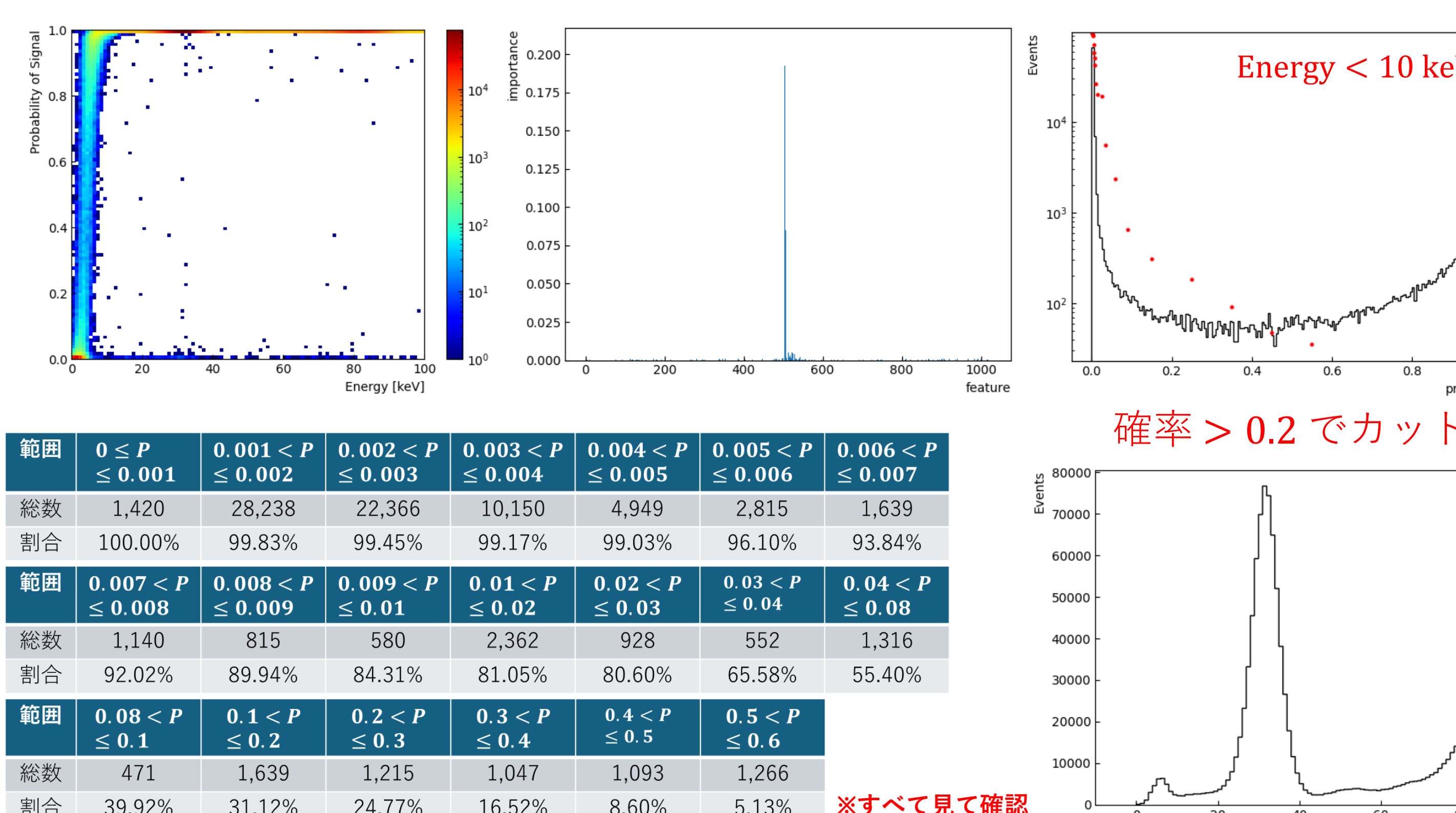
▶ モデルの評価



Log Lossの評価



▶ 機械学習の結果



範囲	$0 \leq P \leq 0.001$	$0.001 < P \leq 0.002$	$0.002 < P \leq 0.003$	$0.003 < P \leq 0.004$	$0.004 < P \leq 0.005$	$0.005 < P \leq 0.006$	$0.006 < P \leq 0.007$
総数	1,420	28,238	22,366	10,150	4,949	2,815	1,639
割合	100.00%	99.83%	99.45%	99.17%	99.03%	96.10%	93.84%
範囲	$0.007 < P \leq 0.008$	$0.008 < P \leq 0.009$	$0.009 < P \leq 0.01$	$0.01 < P \leq 0.02$	$0.02 < P \leq 0.03$	$0.03 < P \leq 0.04$	$0.04 < P \leq 0.08$
総数	1,140	815	580	2,362	928	552	1,316
割合	92.02%	89.94%	84.31%	81.05%	80.60%	65.58%	55.40%
範囲	$0.08 < P \leq 0.1$	$0.1 < P \leq 0.2$	$0.2 < P \leq 0.3$	$0.3 < P \leq 0.4$	$0.4 < P \leq 0.5$	$0.5 < P \leq 0.6$	
総数	471	1,639	1,215	1,047	1,093	1,266	
割合	39.92%	31.12%	24.77%	16.52%	8.60%	5.13%	

※すべて見て確認

IV. まとめと今後の展望

まとめ

- ・機械学習は従来の方法よりもノイズとシグナルの弁別能がある
- ・ランダムフォレストよりもXGBのほうが弁別能は高い

今後の展望

- ・より精度を向上させられるパラメータの決定, データセットの作成を行う
- ・アルゴリズムを組み合わせたり, 深層学習を行ったりすることでより弁別能が高いモデルを作成する

V. 参考文献

- [1] 伏見賢一, (2021) 『宇宙物理学入門-現代宇宙物理学のAからQ-』第3版 大学教育出版
- [2] scikit learn, 'RandomForestClassifier'
<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
(参照2024-06-26)
- [3] Eric Leung, 'XGBoost Documentation'
<https://xgboost.readthedocs.io/en/stable/#>
(参照2024-06-26)