Modified Layerwise Learning for Data Re-uploading Classifier in High-Energy Physics Event Classification

2021 IEEE International Conference on Quantum Computing and Engineering

Eraraya R. Muten*, Quantum Technology Lab, Institut Teknologi Bandung
Togan T. Yusuf, Ankara University
Andrei V. Tomut, Babes-Bolyai University

Special thanks to:

QUARKS INTERACTIVE

Amazon Braket aws

***Presenter**

Introduction

Algorithms

Experimental Setup

Results

Conclusion & Outlook

Outline

Introduction:

- Background
- Related Work

Algorithms:

- Data Re-uploading Classifier
- Modified Layerwise Learning

Experimental Setup:

- Dataset Introduction
- Training Setup

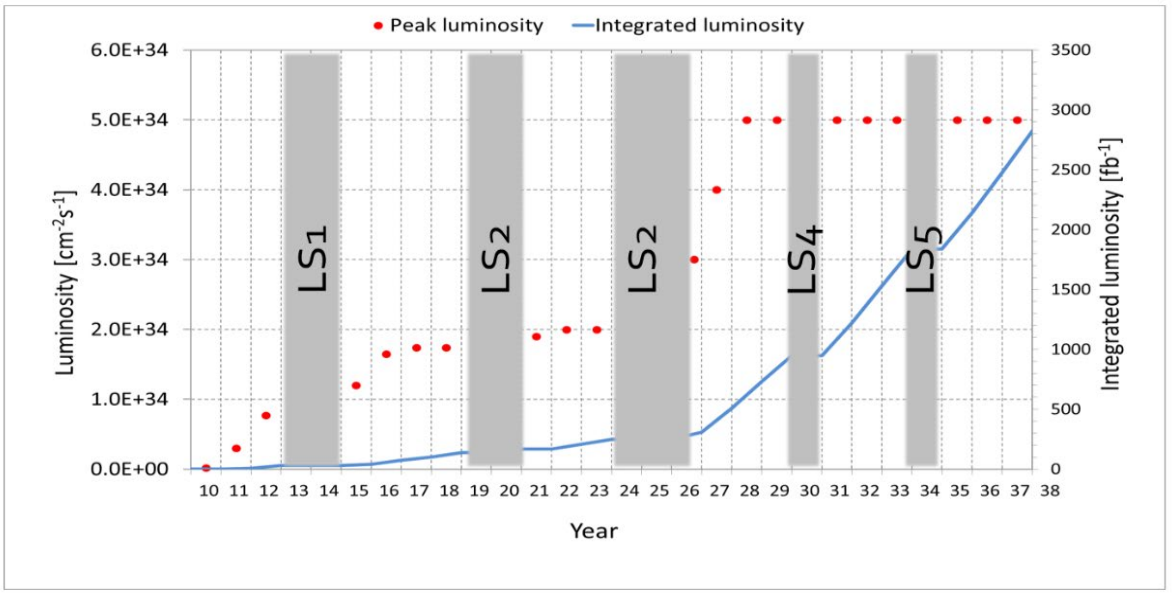
Results

Conclusion & Outlook

Introduction: Background

01

HL-LHC upgrades at CERN will require enormous computing resources[1]



Projected LHC performance through 2038

the amount of data will increase at least 10x

more luminosity = produce more data[1]

[1] Burkhard Schmidt 2016 *J. Phys.: Conf. Ser.* 706 022002.

Introduction: Background

02

Quantum computing has potential in improving performance of data processing and ML[2]

Can it improves HEP simulation and data analysis?

Examples of HEP areas explored:

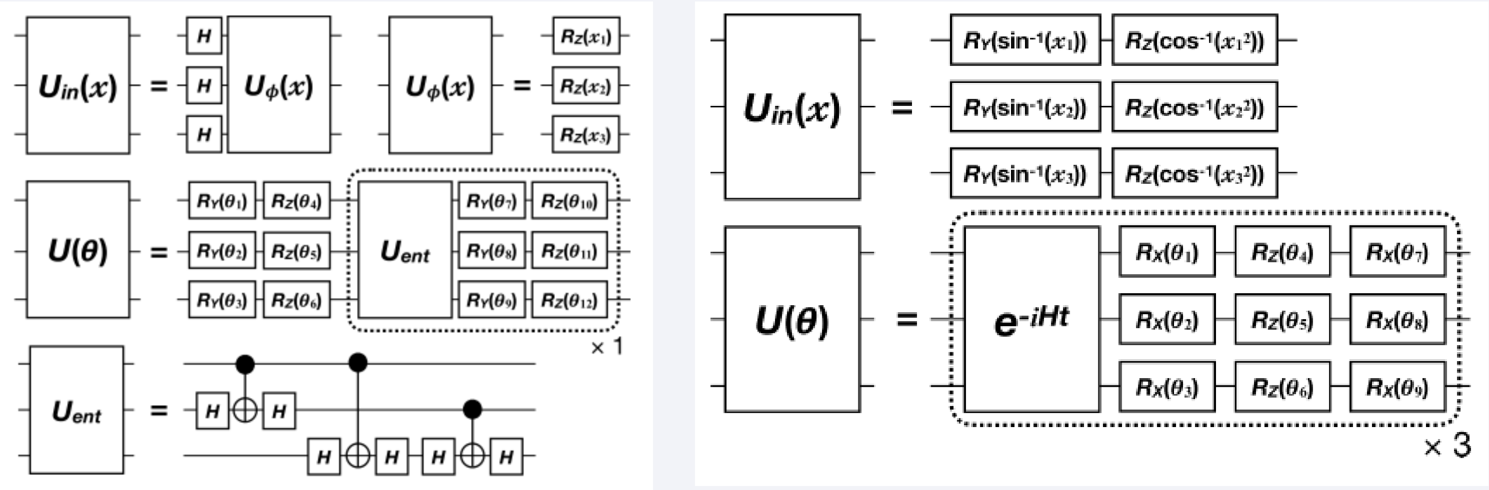
- Higgs optimization problem with quantum annealing[3]
- Identification of charged particle trajectories[4]
- HEP event classification[5, 6]
  Event classification: separate signals from background in the recorded/simulated data.

[2] Biamonte J, et al. *Nature* 2017;549.
[3] A. Mott, et al. *Nature*, vol. 550, no. 7676, pp. 375–379, 2017.
[4] I. Shapoval and P. Calafiura. *EPJ Web of Conferences*, vol. 214, p. 01012, 2019.
[5] J. Chan, et al. *PoS(LeptonPhoton2019)*, vol. 367, 2019, p. 049.
[6] K. Terashi, et al. *Computing and Software for Big Science*, vol. 5, no. 1, p. 2, 2021.

Introduction: Related Work

Related Works

- In [6], a Quantum Support Vector Machine (QSVM)[7] and a Quantum Circuit Learning (QCL)[8] models are trained to classify the SUSY dataset[9]



QSVM and QCL circuits (respectively) used in the study of [6]

[6] K. Terashi, et al. *Computing and Software for Big Science*, vol. 5, no. 1, p. 2, 2021.
[7] V. Havlicek, et al. *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.
[8] K. Mitarai, et al. *Phys. Rev. A*, vol. 98, p. 032309, Sep 2018.
[9] P. Baldi, et al. *Nature Communications*, vol. 5, no. 1, p. 4308, 2014.

Introduction: Related Work

Related Works

- The study showed increasing the number of qubits does not necessarily improve the classifier's performance.
- Both circuits (the QSVM and QCL) employ the angle embedding, which requires one qubit for every feature in the dataset.

The Question

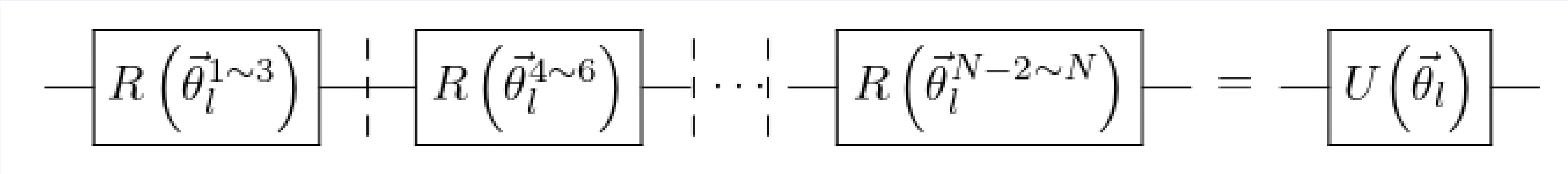
If there is no clear advantage of increasing the number of qubits, how about training one that use very small number of qubits?

Given equal performance, training a model with fewer qubits is both timely and economically more efficient.

[6] K. Terashi, et al. *Computing and Software for Big Science*, vol. 5, no. 1, p. 2, 2021.
[7] V. Havlicek, et al. *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.
[8] K. Mitarai, et al. *Phys. Rev. A*, vol. 98, p. 032309, Sep 2018.
[9] P. Baldi, et al. *Nature Communications*, vol. 5, no. 1, p. 4308, 2014.

Algorithms: Data Re-uploading Classifier (DRC)

It is proven that a single qubit is sufficient to perform universal classification[10]. The authors called it as a data re-uploading classifier.



Circuit schematic of a one qubit DRC's layer

$$R(\tau,\phi,\omega) = \begin{bmatrix} e^{-i\frac{\tau+\omega}{2}}\cos\left(\frac{\phi}{2}\right) & -e^{i\frac{\tau-\omega}{2}}\sin\left(\frac{\phi}{2}\right) \\ e^{-i\frac{\tau-\omega}{2}}\sin\left(\frac{\phi}{2}\right) & e^{i\frac{\tau+\omega}{2}}\cos\left(\frac{\phi}{2}\right) \end{bmatrix}$$

$$\vec{\theta}_l^{n\sim n+2} = \left(\theta_l^n, \theta_l^{n+1}, \theta_l^{n+2}\right)$$

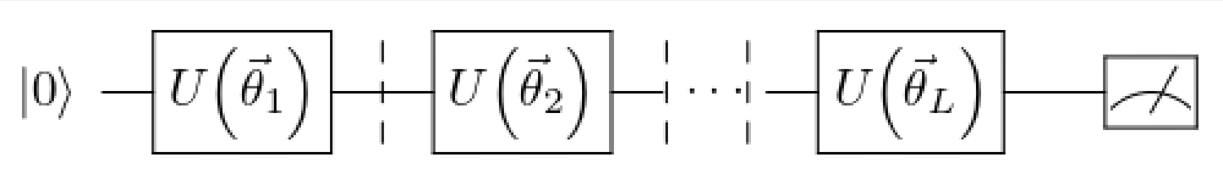
$$\vec{\theta}_l = \left(\theta_l^1, \theta_l^2, \theta_l^3, \ldots, \theta_l^N\right)$$

$$\theta_l^n = w_l^n x^n + b_l^n$$

One main advantage of DRC: in theory, the number of required qubits is independent of the number of features.

[10] A. Perez-Salinas, et al. *Quantum*, vol. 4, p. 226, Feb. 2020.

Algorithms: Data Re-uploading Classifier (DRC)



A complete one qubit DRC circuit is the repetition of the layer followed by a measurement

If we set background = |0⟩ and signal = |1⟩, the classification task now is equivalent to maximizing the fidelity between the output quantum state with the respective quantum state label.

$$J(\vec{\alpha},\vec{\theta}) = \frac{1}{2M}\sum_{m=1}^{M} \text{sum}\left\{\left(\vec{y}_{\text{pred}_m}(\vec{\alpha},\vec{\theta}) - \vec{y}_{\text{true }m}\right)^2\right\}$$

$$\vec{y}_{\text{pred}_m}(\vec{\alpha},\vec{\theta}) = \vec{\alpha}\odot\begin{bmatrix}\left\langle O_0(\vec{\theta})\right\rangle_m \\ \left\langle O_1(\vec{\theta})\right\rangle_m\end{bmatrix}$$

$$\left\langle O_0(\vec{\theta})\right\rangle_m = {}_m\left\langle\Psi_{DRC}(\vec{\theta})|O_0|\Psi_{DRC}(\vec{\theta})\right\rangle_m$$

$$\left\langle O_1(\vec{\theta})\right\rangle_m = {}_m\left\langle\Psi_{DRC}(\vec{\theta})|O_1|\Psi_{DRC}(\vec{\theta})\right\rangle_m$$

$$\vec{\alpha} = \begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix}$$

$$\left|\Psi_{DRC}(\vec{\theta})\right\rangle_m = U\left(\vec{\theta}_L\right)U\left(\vec{\theta}_{L-1}\right)\dots U\left(\vec{\theta}_1\right)|0\rangle$$

$$O_0 = |0\rangle\langle 0|$$

$$O_1 = |1\rangle\langle 1|$$

Algorithms: Modified Layerwise Learning

Layerwise learning is a training strategy that trains only subset of parameters at a time, ensuring a favorable signal-to-noise ratio[11].

Help avoid the problem of barren plateaus thanks to:
- low circuit's depth
- low number of parameters optimized in one update step
- larger gradients magnitude

We trained the parameter of each circuit layer one at a time (freezing the parameters of the other layers) once, and trained the whole circuit once.

[11] A. Skolik, et al. *Quantum Machine Intelligence*, vol. 3, no. 1, p. 5, 2021.

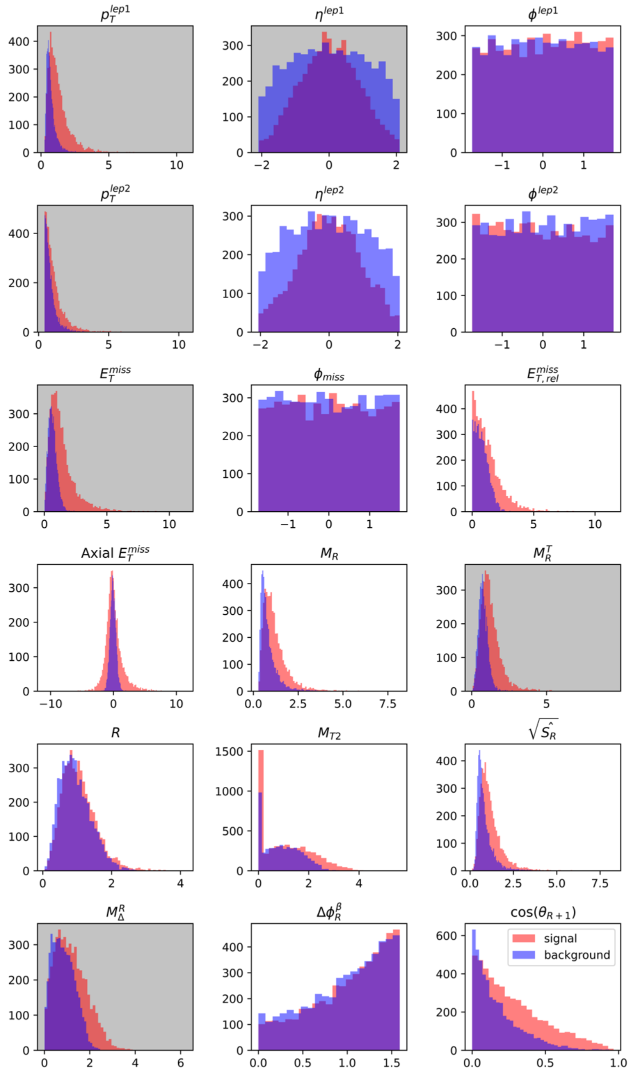
Experimental Setup: The Dataset

We chosed SUSY dataset[9], the one also studied in [6]

- Signal/true label: a chargino-pair production via the Higgs boson and a W-boson
- Background: W-boson pair production

Both processes have the same final state, a charged lepton and a neutrino from the decayed W-boson. The chargino-pair decay into a neutralino that avoids detection.

[6] K. Terashi, et al. *Computing and Software for Big Science*, vol. 5, no. 1, p. 2, 2021.
[9] P. Baldi, et al. *Nature Communications*, vol. 5, no. 1, p. 4308, 2014.



Experimental Setup: The Dataset

Entire dataset includes about 5 million events, we used 10,000 samples from it.

Each signal is characterized by 18 features:

- The first 8 features are kinematic properties (transverse momentum P_T, *pseudo-rapidity* η, azimuthal angle ϕ, energy E_T)
- The rest of them are derived from (functions of) the first 8.

Among 18, we selected: $p_{\mathrm{T}}^{\mathrm{lep1}}, p_{\mathrm{T}}^{\mathrm{lep2}}, E_T^{\mathrm{miss}}, M_R^T, M_\Delta^R, \eta^{\mathrm{lep1}}$

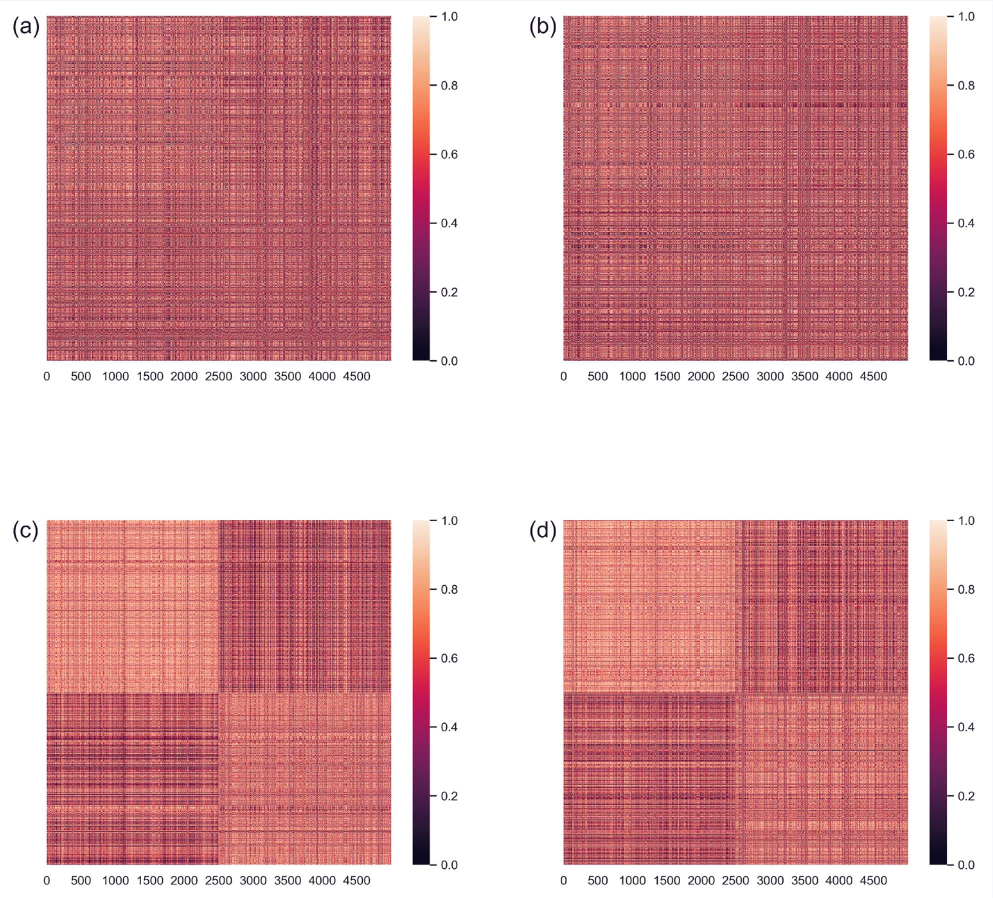
With 6 features, no zero padding is needed.

Experimental Setup: Training Setup

- Trained on the PennyLane[12] state-vector simulator
- The number of layers of the DRC in this study is 5 (62 trainable parameters)
- 10 epochs/training with batch size of 128 samples
- Parameter optimization by Adam[13] optimizer with 0.05 learning rate
- Performance metric: AUC (area under ROC curve) value
- After training, we tested the model on Rigetti's quantum processor Aspen-9 through Amazon Braket for 2000 samples

[12] V. Bergholm, et al. arXiv:1811.04968.
[13] D. P. Kingma and J. Ba. *ICLR* 2015.

Results



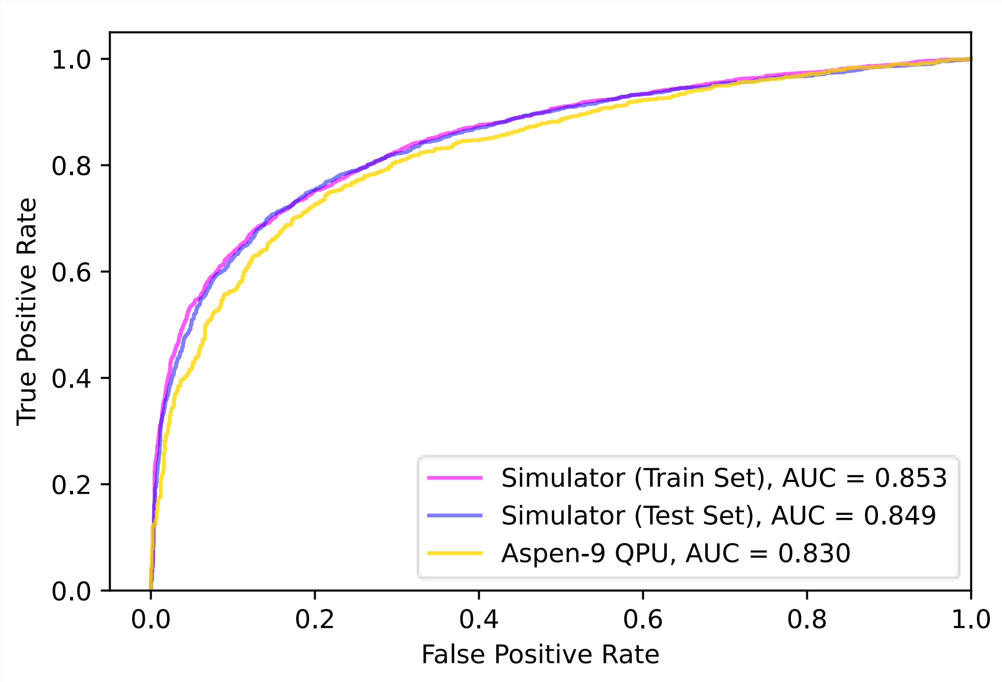
Top row: before training
Bottom row: after training

Left column: train set
Right column: test set

The classifier was able to differentiate between classes after the training.

$$F_{i,j} = \left| {}_i\left\langle \Psi_{DRC}(\vec{\theta}) \mid \Psi_{DRC}(\vec{\theta}) \right\rangle_j \right|^2$$

Results



ROC Curves and AUC value of the classifier after the training.

The classifier was able to generalize well.
Agreeing with the study of [6], running the classifier on QPU may lead to worse performance due to errors from noisy hardware.

Results

AUC VALUE COMPARISON

	Backend	AUC
QSVM [1]	Johannesburg QPU, IBM Q (3-qubits circuit)	0.799 ± 0.020
	Boeblingen QPU, IBM Q (3-qubits circuit)	0.807 ± 0.010
	QASM simulator (3-qubits circuit)	0.815 ± 0.015
QCL [1]	Qulacs simulator (3-qubits circuit)	0.833 ± 0.063
DRC	PennyLane simulator (1-qubit circuit)	0.849
	Rigetti's Aspen-9 QPU, AWS (1-qubit circuit)	0.830

DRC used fewest number of qubit but better: increasing the number of qubits does not always result in better performance.

Other important factors: embed the classical data to the circuit, the structure of the circuit, and how to train the circuit hold an equally important role.

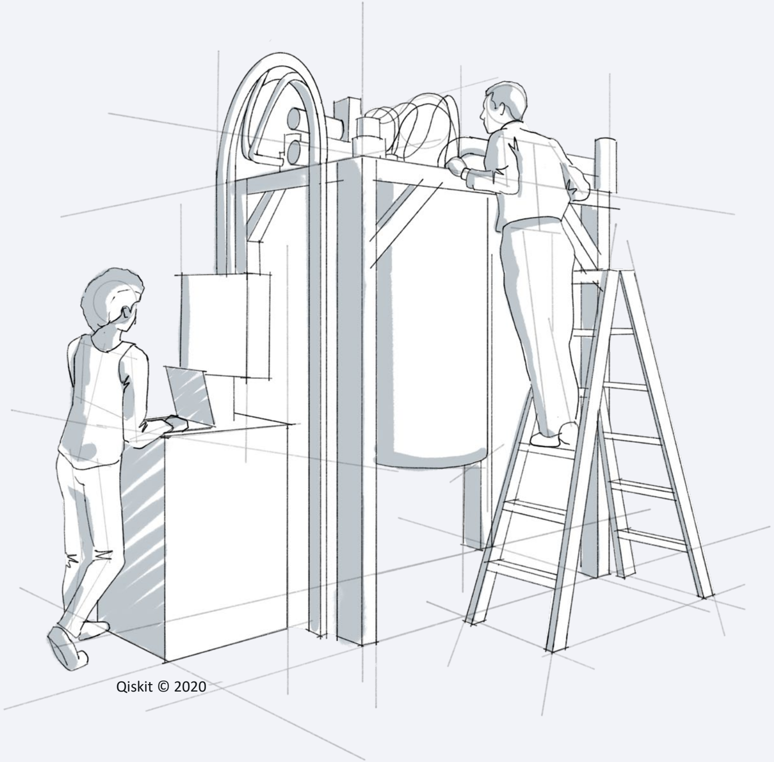
Conclusion

- Data re-uploading classifier with one qubit, trained with the modified layerwise learning, is able to perform better than the compared methods on event classification of the SUSY dataset.
- The AUC value obtained from the simulator is also close to the one obtained from running the test on the quantum hardware.
- A promising approach for future research in HEP with larger datasets since it requires fewer qubits, leading to less queue time and computational power required.

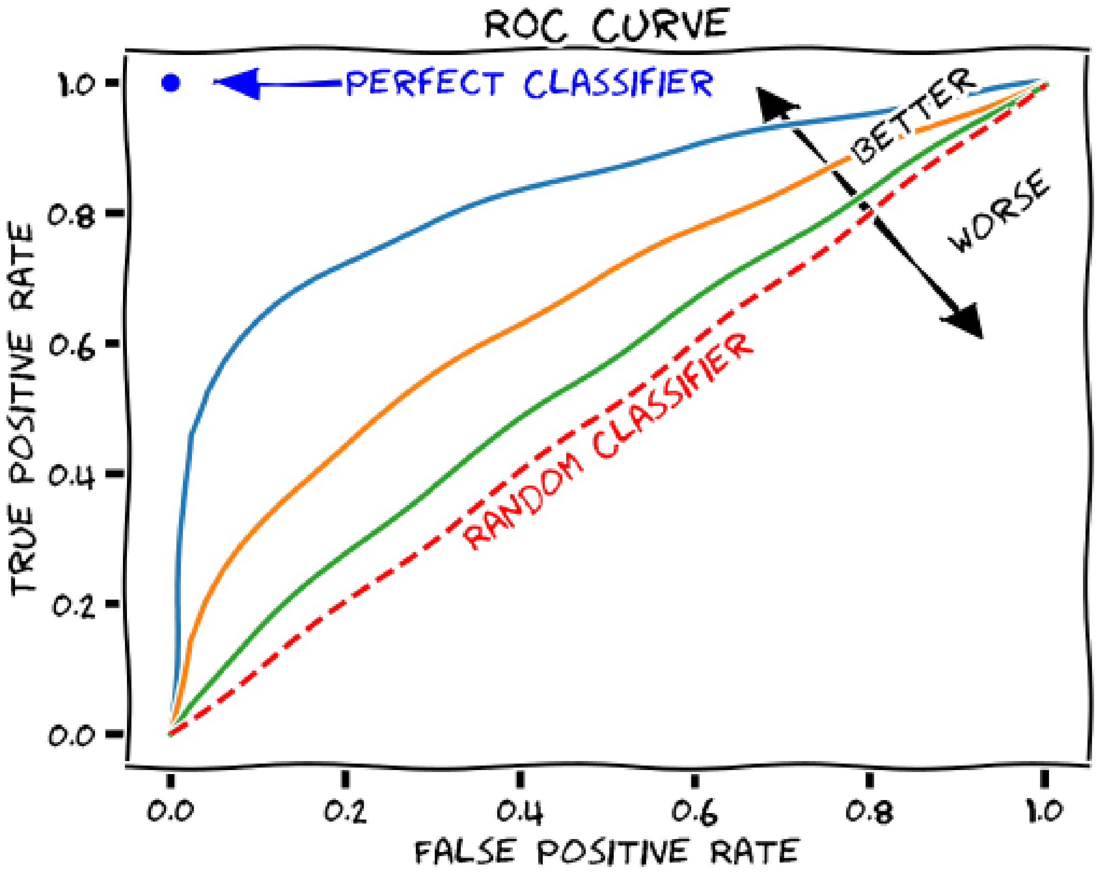
Outlook

- Train directly on quantum hardware > taking noise into account during the training?
- DRC can be expanded to multi-qubits version, how does increasing the number of qubits in DRC affect the performance?
- How the model perform on larger scale of dataset (> 1 million samples)?

Thank You!
Any Questions?



APPENDIX


ROC CURVE
PERFECT CLASSIFIER
BETTER
WORSE
RANDOM CLASSIFIER
TRUE POSITIVE RATE
FALSE POSITIVE RATE
0.0
0.2
0.4
0.6
0.8
1.0