Real-Time Detection of Charge Jumps in Superconducting Qubits with a Convolutional Neural Network

Daniel Gaytan-Villarreal*
Carnegie Mellon University

Peter Meiring
Carnegie Mellon University

Daniel Baxter
Fermilab

Daniel Bowring
Fermilab

Grace Bratrud
Fermilab

Matteo Cremonesi
Carnegie Mellon University

Giuseppe Di Guglielmo
Fermilab

Grace Wagner
Colorado School of Mines

Bowen Xiao
Carnegie Mellon University

Abstract

Ionizing radiation from cosmic rays and ambient radioactivity induces charge jumps in superconducting qubits, causing correlated errors that challenge fault-tolerant quantum computing while simultaneously providing a detection signature for quantum sensing applications. Current detection methods operate offline, providing only a post-hoc record of charge events. By the time a jump is identified, the affected experiment is already over, and no adaptive response is possible. In this work, a real-time charge-jump detector for superconducting qubits, based on a dilated causal convolutional neural network (DC-CNN) designed for in-the-loop deployment on the Quantum Instrumentation Control Kit (QICK) platform, is presented. The network is trained on synthetic Ramsey tomography scans generated from qubit templates measured at the Northwestern Experimental Underground Site (NEXUS) at Fermilab, and translated to FPGA firmware via HLS4ML with optimized fixed-point quantization, achieving a low, microsecond-scale per-inference latency on the Zynq UltraScale+ RFSoC ZCU216. At this operating point, the DC-CNN matches the detection efficiency of the established offline χ^2 algorithm across a broad range of charge-jump magnitudes at a matched false-positive rate, while requiring no per-qubit hyperparameter tuning. This shifts charge-jump detection from a post-hoc diagnostic to a control-loop primitive, enabling adaptive protocols that respond to radiation-induced events in situ, with applications to quantum-computing error mitigation and to the use of superconducting qubits as particle detectors.

1 Introduction

Superconducting transmon qubits operate as artificial atoms whose sensitivity to charge noise is set by the ratio E_J/E_C of Josephson to charging energy. Ionizing radiation incident on the qubit substrate generates phonon bursts that break Cooper pairs and produce quasiparticles, manifesting as sudden discontinuous shifts in the offset charge sensed by the qubit island [Cardani et al., 2023]. These charge jumps are spatially correlated across millimeter-scale chips and accompanied by transient suppression of relaxation times [Wilen et al., 2021], posing a fundamental challenge to error-correction schemes that assume uncorrelated errors. The same sensitivity that makes qubits vulnerable to ionizing

*Corresponding author: jgaytanv@andrew.cmu.edu

Preprint.

radiation also enables their use as sensors for sub-GeV dark matter and other beyond-Standard-Model phenomena [Linehan et al., 2025].

Measurements at NEXUS, located 107 m underground at Fermilab with a movable lead shield providing 4π gamma coverage, employ Ramsey tomography (X/2–Idle–X/2) on transmons operated in a mildly charge-sensitive regime ($E_J/E_C = 24$) to make charge jumps directly observable [Bratrud et al., 2025]. The conventional offline analysis fits each N-sample charge sweep to a qubit-specific template $\hat{x}(\theta)$ and monitors a rolling reduced χ^2 statistic over prefixes of length $n \in [1, N]$:

$$\chi^2_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \hat{x}(\theta))^2}{\hat{\sigma}^2(\theta)}, \qquad \theta_n^{\min} = \arg\min_{\theta} \chi^2_n(\theta). \tag{1}$$

A charge jump within the scan introduces a discontinuity in $\theta_n^{\min}$ that drives $\chi^2_n(\theta_n^{\min})$ to rise rapidly with each subsequent sample; a jump is tagged when this rolling value crosses a per-qubit threshold, and the procedure then resets with the first above-threshold sample becoming $i = 1$ in the next iteration. While effective offline, this approach requires per-qubit threshold tuning, has reduced sensitivity to jumps occurring within the first $\sim$20 samples of a scan (where insufficient pre-jump data has accumulated for the rolling statistic to build up), and imposes a non-trivial post-acquisition latency that precludes real-time response.

In this work we recast charge-jump detection as a binary classification problem amenable to fixed-latency neural-network inference. We train a Dilated Causal CNN [Bai et al., 2018] on synthetic Ramsey scans generated from templates measured at NEXUS, quantize it to fixed-point precision, and compile it to FPGA firmware via HLS4ML [Duarte et al., 2018]. The resulting network matches the offline χ^2 baseline at matched false-positive rate without per-qubit tuning, and runs at fixed latency well within a single experiment cycle on the QICK platform [Stefanazzi et al., 2022, Ding et al., 2024, Di Guglielmo et al., 2025].

2 Method

Architecture. We compare a Temporal Convolutional Network (TCN) against a Dilated Causal CNN (DCCNN) obtained by ablating the residual paths from the TCN. Both share the same four 1D convolution layers (8 filters, kernel 3, dilation 1 in block 1; 16 filters, dilation 2 in block 2), depth, and BatchNorm/ReLU stacks; only the 1×1 skip projections differ (Fig. 1). Both the DCCNN and TCN feature lightweight parameter footprints. Inputs are 80×2 tensors formed by concatenating each Ramsey scan with its first-difference channel $\Delta x_i = x_i - x_{i-1}$, which exposes the discontinuous phase shift produced by a jump.

Data. We generate synthetic Ramsey scans following the procedure of Bratrud et al. [2025]: per-qubit templates $\hat{x}(\theta)$ and noise envelopes $\hat{\sigma}(\theta)$ for five configurations (Q1–Q3 superconducting, Q4 in superconducting and shield-open modes) are taken from NEXUS calibration data. Jumps are drawn at a characteristic mHz-scale rate with magnitudes uniform in $[0.01, 0.5]\,e$ and random sign; per-scan phases are time-ordered so the post-event phase of scan k becomes the pre-event phase of scan $k+1$. Gaussian noise is added using the qubit-specific $\hat{\sigma}$. A substantial fraction of the generated scans contain at least one charge jump.

Training and quantization. We split the scans for train/val/test with stratified sampling, and train with Adam and binary cross-entropy on a single A100 GPU using Keras's deterministic mode (seed 42, early stopping on val loss). The float model is then quantized via QKeras [Coelho et al., 2021, 2019] with `Conv1D`→`QConv1D` substitution and global pool replaced by an HLS4ML-compatible `AvgPool`. A bit-width sweep over standard fixed-point precisions confirmed that the network meets the strict FPGA resource and latency budget while preserving high accuracy.

3 Results

The DCCNN outperforms the TCN on every per-class binary classification metric (precision, recall, and F1 score) in both the float and quantized regimes. This suggests that the residual paths add parameters without providing a detection benefit for this particular time-series classification task.

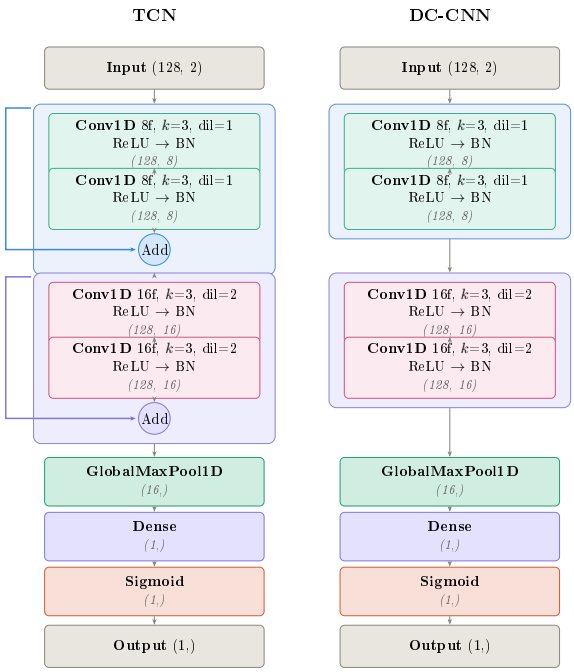

Figure 1: Proposed architectures: TCN (left) and DCCNN (right). The DCCNN has the same layer structure as the TCN minus the residual connections. Tensor shapes are annotated inside each block.

While the No-Jump class yields high recall across both architectures, Jump recall remains the limiting factor. Consequently, we select the DCCNN for FPGA deployment.

The key evaluation for downstream physics applications is against the established χ^2-window method. To match the methodology of Bratrud et al. [2025], we evaluate both methods on simulated scans per qubit with jumps uniform in $[0.01, 0.5]\,e$. The χ^2 threshold is tuned per qubit on the training set to maximize jump-tagging efficiency at a baseline false-positive rate (FPR) matched to the DCCNN's post-quantization performance. The quantized DCCNN achieves comparable mean per-qubit detection efficiency to the per-qubit-tuned χ^2 across all five templates. The DCCNN accomplishes this using a single set of weights for all qubits, eliminating the need for individual calibration.

4 Firmware Deployment

The Quantum Instrumentation Control Kit (QICK) [Stefanazzi et al., 2022, Ding et al., 2024, Fermi National Accelerator Laboratory, 2024] is an open-source qubit control platform built on Xilinx RFSoC FPGAs, providing a unified framework for superconducting qubit control and readout. Running on the ZCU216 evaluation board, QICK delivers 16 DAC channels at 9.85 GS/s and 16 ADC channels at 2.5 GS/s, supports direct mixer-free signal generation up to 10 GHz, and coordinates pulse sequences through a timed processor (tProc) with 100 ps timing resolution. Recent work has demonstrated that custom neural-network IP blocks compiled via HLS4ML can be integrated directly into the QICK readout chain for low-latency qubit-state discrimination [Di Guglielmo et al., 2025]; we adapt the same workflow to charge-jump detection.

We compile the quantized DCCNN through HLS4ML into a standalone IP block and integrate it into the QICK firmware on the Zynq UltraScale+ RFSoC ZCU216 (Vivado 2023.1). Time-series data from the qubit readout chain feed the IP via an AXI4-Stream interface; the binary jump flag is exposed via an AXI4-Lite register, which the QICK tProc polls and can use to dynamically branch its instruction sequence in real time.

Post-place-and-route resource utilization confirms that the DCCNN IP block integrates alongside the baseline QICK infrastructure. It occupies a modest fraction of available flip-flops (FF), look-up tables

(LUT), and digital signal processors (DSP), with block RAM (BRAM 36K) acting as the primary constraint. Vitis HLS C-synthesis at a 3.33 ns target clock estimates a fixed, microsecond-scale per-inference latency and a tight initiation interval, ensuring sustained high throughput well within the required bounds of standard experimental readout cycles.

5 Conclusion

We have presented a real-time, FPGA-deployable charge-jump detector for superconducting qubits based on a compact dilated causal CNN quantized to fixed-point precision. The network achieves a mean per-qubit detection efficiency on $|\Delta q| \in [0.1, 0.5]\, e$ comparable to the per-qubit-tuned χ^2 baseline at a matched false-positive rate. Unlike the baseline, it runs as a stateless, fixed-latency forward pass that integrates directly into the QICK control loop. A single trained model successfully serves all five qubit configurations, eliminating the per-qubit calibration required by the χ^2 method. These results show that ML-based charge-jump detection is viable under the resource and latency constraints of on-chip quantum firmware, opening a path to in-the-loop adaptive protocols and triggering systems that respond to radiation-induced quantum events as they occur.

Acknowledgments and Disclosure of Funding

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. The authors thank the NEXUS collaboration for providing access to the qubit calibration data and explaining their jump-finding methodology.

References

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018. URL `http://arxiv.org/abs/1803.01271`.

G. Bratrud et al. Measurement of correlated charge noise in superconducting qubits at an underground facility. *Nature Commun.*, 16(1):9906, 2025. doi: 10.1038/s41467-025-63724-4.

L. Cardani et al. Disentangling the sources of ionizing radiation in superconducting qubits. *Eur. Phys. J. C*, 83 (1):94, 2023. doi: 10.1140/epjc/s10052-023-11199-2.

Claudionor N. Coelho, Aki Kuusela, Shan Li, Hao Zhuang, Jennifer Ngadiuba, Thea Klaeboe Aarrestad, Vladimir Loncar, Maurizio Pierini, Adrian Alan Pol, and Sioni Summers. Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors. *Nature Machine Intelligence*, 3(8):675–686, 2021. doi: 10.1038/s42256-021-00356-5.

Claudionor N. Coelho et al. QKeras: a quantization deep learning library for Tensorflow Keras. `https://github.com/google/qkeras`, 2019.

Giuseppe Di Guglielmo et al. End-to-End Workflow for Machine-Learning-Based Qubit Readout With QICK and hls4ml. *IEEE Trans. Quantum Eng.*, 6:1–10, 2025. doi: 10.1109/TQE.2025.3604712.

Chunyang Ding et al. Experimental advances with the QICK (Quantum Instrumentation Control Kit) for superconducting quantum hardware. *Phys. Rev. Res.*, 6(1):013305, 2024. doi: 10.1103/PhysRevResearch.6.013305.

Javier Duarte et al. Fast inference of deep neural networks in FPGAs for particle physics. *JINST*, 13(07):P07027, 2018. doi: 10.1088/1748-0221/13/07/P07027.

Fermi National Accelerator Laboratory. QICK: Quantum Instrumentation Control Kit. `https://github.com/openquantumhardware/qick`, 2024. Open-source software, accessed [insert date].

R. Linehan et al. Estimating the energy threshold of phonon-mediated superconducting qubit detectors operated in an energy-relaxation sensing scheme. *Phys. Rev. D*, 111(6):063047, 2025. doi: 10.1103/PhysRevD.111.063047.

Leandro Stefanazzi et al. The QICK (Quantum Instrumentation Control Kit): Readout and control for qubits and detectors. *Rev. Sci. Instrum.*, 93(4):044709, 2022. doi: 10.1063/5.0076249.

C. D. Wilen et al. Correlated charge noise and relaxation errors in superconducting qubits. *Nature*, 594(7863): 369–373, 2021. doi: 10.1038/s41586-021-03557-5.