Development of a MEMS-based Earthquake Dataset using the Raspberry Shake Network in New Zealand

Tasfia Zaman Samiha
Department of Computer Science and Engineering
Khulna University of Engineering and Technology
Khulna, Bangladesh

Danuka Ravishan
Joint Centre for Disaster Research
Massey University
Wellington, New Zealand

***Abstract*—Traditional seismic monitoring is often limited by the high cost of instrumentation and logistical barriers, hindering the expansion of earthquake monitoring networks especially in under-resourced regions. Low-cost MEMS-based sensors offer a scalable alternative, but require specialized datasets to train machine learning models adapted to their unique noise characteristics and sensitivity profiles. To address this, we systematically collected waveforms from approximately 4,000 earthquakes (magnitude 2.7 to the highest recorded) recorded across 89 Raspberry Shake stations in New Zealand from 2020–2025. Events were matched to nearby stations based on epicentral distance criteria (100 km for M 2.7–5.5, 150 km for M>5.5). A staged filtering pipeline using PhaseNet, EQTransformer, and GPD models, cross-validated with theoretical TauP arrivals, was applied to ensure phase pick quality across three confidence tiers. The final curated dataset comprises 918 high-confidence waveforms with validated P and S wave arrivals, alongside approximately 16,035 total waveform records spanning all quality tiers. This dataset addresses the critical scarcity of labeled training data for low-cost seismic instrumentation, enabling the development of phase pickers specifically calibrated for MEMS sensors.**

***Index Terms*—MEMS, seismic dataset, Raspberry Shake, earthquake monitoring, machine learning, phase picking, New Zealand**

I. Introduction

Earthquake research increasingly relies on high-quality seismic datasets. Established databases such as STEAD [1] and INSTANCE [2] provide comprehensive collections from traditional seismometers, but are limited to high-cost instruments. Additional regional benchmark datasets including DiTing [3] for China and the curated Pacific Northwest (PNW) dataset [16] have further demonstrated that geographically and instrumentally diverse training corpora are essential for robust machine learning phase pickers. Low-cost MEMS accelerometers have emerged as promising alternatives for seismic monitoring [8]–[11], offering affordability and ease of deployment for large-scale networks. The Raspberry Shake network exemplifies this approach, employing MEMS-based sensors for real-time seismic monitoring globally [17].

Deep learning has become the dominant paradigm for seismic signal processing [1], [6], yet models trained on traditional broadband data often generalize poorly to MEMS recordings, underscoring the need for sensor-specific training corpora. In New Zealand, CRISiSLab maintains a dense sensor distribution across Wellington and surrounding areas. A recent study demonstrated that Raspberry Shake data can augment the national GeoNet network to improve local earthquake location and characterization [13], illustrating the potential of citizen-science seismometers for operational monitoring.

Despite growing adoption of MEMS sensors [12], a critical gap remains: the lack of publicly available, high-quality MEMS-based earthquake datasets. New Zealand's frequent seismic activity, monitored by the comprehensive GeoNet network [13], and driven by its position at the boundary of the Pacific and Australian tectonic plates including the active Hikurangi subduction zone and Alpine Fault system makes it an ideal region for developing such resources, as most of the Raspberry Shake sensors are distributed along the Alpine Fault line. Hence we fill this gap by generating a comprehensive MEMS-based earthquake dataset from the Raspberry Shake network.

II. Literature Review

A. Benchmark Seismic Datasets for Machine Learning

Large labeled seismic datasets have been the primary driver of machine learning progress in seismology. STEAD [1], the most widely used global benchmark, contains over one million broadband three-component waveforms with annotated P and S picks. INSTANCE [2] extended this to Italy with $\sim$1.2 million waveforms from 54,000 events across 620 stations. DiTing [3] contributed 2.7 million traces from Chinese regional events (M 0–7.7), while the Pacific Northwest (PNW) dataset [16] provides a 21-year AI-ready collection of $\sim$190,000 waveforms. All four benchmarks rely exclusively on high-cost broadband or strong-motion instruments, leaving a critical gap for low-cost MEMS sensor data.

B. Deep Learning Phase Pickers and MEMS Sensors

PhaseNet [6] introduced a U-Net architecture producing per-sample P, S, and noise probability distributions, achieving near-expert analyst performance. EQTransformer [4] employs hierarchical attention for simultaneous detection and picking, demonstrating superior event recovery on single-station data. GPD [5] trained convolutional networks on millions of Southern California waveforms, yielding a body-wave detector that is robust to high background noise. These three models are used here as complementary cross-validation tools. On the sensor side, low-cost MEMS accelerometers have been

https://doi.org/10.54389/EJDX1441

extensively reviewed for seismological use [8]; MEMS arrays have been shown to enhance early warning [11], and labeled MEMS datasets from Italian seismic sequences have confirmed the scientific value of low-cost recordings even under elevated noise [9], [10]. The Raspberry Shake network, the largest citizen-science seismic network globally [17], has been validated against national networks in Bucharest [14] and New Zealand [13], establishing that systematic quality control can produce research-grade MEMS waveforms suitable for machine learning development.

III. Data Collection and Processing

A. Event and Station Selection

Earthquake events were sourced from GeoNet's Quake Search service [13] covering January 2019 to June 2025 with a minimum magnitude threshold of M 2.7, chosen to ensure detectable signal amplitudes on MEMS accelerometers at local distances. The lower bound was set empirically: below M 2.7, signal-to-noise ratios on MEMS channels degraded to levels where reliable automated phase detection became unreliable across the majority of stations. This query returned 4,274 events with focal depths ranging from near-surface to 538 km and magnitudes spanning 2.7–6.0.

Seismic station metadata were retrieved from the Raspberry Shake StationView portal [17]. From an initial candidate pool of 67 New Zealand stations, 44 were confirmed active with the requisite four-channel configuration (EHZ, ENE, ENN, ENZ). A final set of 43 stations entered the waveform acquisition pipeline after one station was excluded due to persistent data gaps throughout the study period. Station verification was performed programmatically via the FDSN web service using ObsPy [15]: each candidate station was queried for its available channel inventory, and only those providing all three MEMS accelerometer channels (ENE, ENN, ENZ) alongside the vertical geophone (EHZ) were retained.

To manage computational load and focus on high-quality recordings, epicentral distance thresholds were applied before waveform download. Earthquake–station pairs with epicentral distances exceeding 100 km for M 2.7–5.5 events, or 150 km for M>5.5 events, were excluded. These thresholds balance signal-to-noise (SNR) considerations where larger distances produce attenuated, noisier waveforms on MEMS channels with the goal of maintaining adequate spatial coverage across the New Zealand network.

B. Waveform Acquisition and Preprocessing

Waveforms were acquired via the FDSN web service using ObsPy [15] in MiniSEED format. Each trace was downloaded as a 180-second window comprising 60 s of pre-event noise and 120 s of post-event signal, centered on the reported GeoNet origin time. This window length was selected to capture the full P-to-S wave sequence for events at the maximum permitted epicentral distances while retaining a statistically robust pre-event noise segment for SNR computation and noise characterization.

All downloaded waveforms were resampled to a uniform 100 Hz sampling rate and stored as four-component MiniSEED files per earthquake–station pair (ENE, ENN, ENZ, EHZ). A detrend and demean step was applied to each trace to remove DC offset and long-period drift before phase detection. Instrument response removal was not applied at this stage, as the downstream deep learning phase pickers were trained on raw count data and the focus of this dataset is phase arrival timing rather than ground-motion amplitude calibration.

C. Five-Stage Hierarchical Quality Control Pipeline

The core of the processing workflow is a five-stage hierarchical quality control (QC) pipeline that combines three complementary deep learning phase pickers PhaseNet [6], EQTransformer [4], and GPD [5] with theoretical travel-time predictions computed using the TauP toolkit [7] under the AK135 velocity model. For each waveform, the expected P- and S-wave arrival times relative to event origin were computed from the reported hypocentral parameters, providing an independent reference against which picker outputs could be cross-validated.

Stage 1 retained waveforms where PhaseNet reported a P-phase detection probability $\geq$0.9, and the PhaseNet P pick fell within $\pm$1 s of the TauP AK135 theoretical arrival. This stage applied the most stringent simultaneous constraints on picker confidence and travel-time consistency, yielding 553 waveforms.

Stage 2 relaxed the PhaseNet probability threshold to the range 0.5–0.9 while extending the cross-validation tolerance to $\pm$2 s. Waveforms passing Stage 2 are those where PhaseNet had moderate confidence and the pick remained temporally consistent with theoretical predictions, contributing 1,505 additional records.

Stage 3 shifted the primary detector to EQTransformer, accepting picks with probability $\geq$0.8 cross-validated against both GPD detections and AK135 theoretical arrivals within $\pm$2 s. This iterative approach captured events where EQTransformer's attention-based architecture outperformed PhaseNet, adding 701 waveforms.

Stages 4 and 5 applied analogous iterative consensus schemes using GPD as the primary detector and a broadened PhaseNet window (0.5–0.8 probability), respectively, each requiring multi-model agreement within $\pm$2 s. These later stages targeted the most challenging low-SNR recordings where single-model confidence was inherently reduced.

Records failing all five stages were flagged as low-quality and retained as noise examples in the dataset rather than discarded, as unvalidated records dominated by anthropogenic or environmental noise are scientifically valuable for noise characterization and augmentation-based training [9]. The spatial distribution of events and stations is illustrated in Fig. 1.

IV. Dataset Characteristics and Validation

A. Dataset Composition and Metadata

The final dataset comprises 16,035 four-component waveform records. Each record contains the vertical geophone chan-

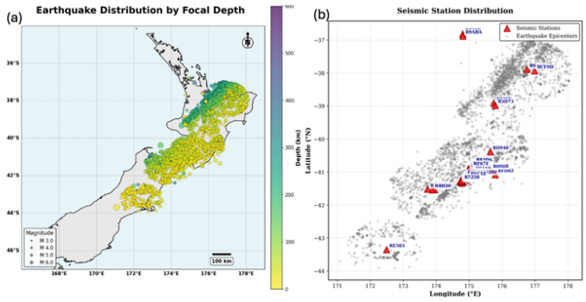

Fig. 1: Spatial distribution of seismicity and monitoring infrastructure. (a) Seismic event distribution colored by focal depth (yellow=shallow, purple=deep, 0–600 km) and scaled by magnitude (M 3.0–6.0), revealing the Hikurangi subduction zone, Alpine Fault system, and deep intraslab events. Scale bar: 100 km. (b) Raspberry Shake stations (red triangles) and earthquake epicentres (gray circles) within 171°–178°E, 37°–44°S.

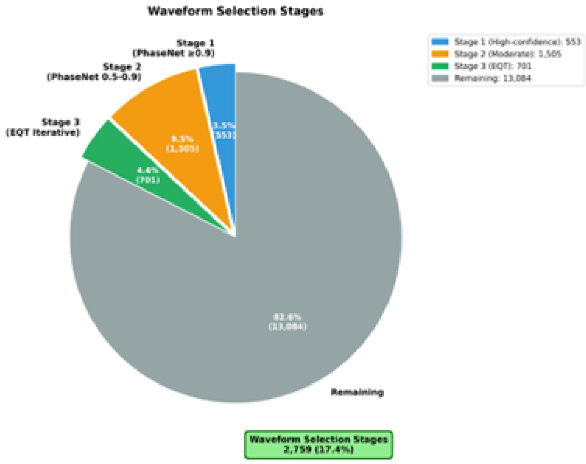

Fig. 2: Proportion of waveforms passing each validation stage. Of 15,843 processed records, 17.4% satisfied cross-validation criteria across multiple phase-picking algorithms; the remainder are unvalidated noise-dominated records retained for noise characterization.

nel (EHZ) and three MEMS accelerometer channels (ENE, ENN, ENZ); where both ENZ and EHZ were available, the selection was based on data continuity. Comprehensive metadata accompany every record, including event geographic coordinates, origin time, magnitude, and focal depth from GeoNet; station identifier, network code, and nominal coordinates from the Raspberry Shake StationView; and derived path parameters epicentral distance, hypocentral distance, and back azimuth computed from the event–station geometry.

The magnitude distribution of the full dataset follows the natural Gutenberg–Richter frequency–magnitude relationship, with the majority of events in the M 2.7–4.0 range and progressively fewer events at larger magnitudes up to M 6.0. This natural distribution reflects the completeness of the GeoNet catalog above M 2.7 in the study region and ensures that the dataset provides broad coverage across the seismic spectrum rather than being biased toward high-magnitude events. Focal depths span from near-surface crustal events (<10 km) to deep intraslab events at 538 km, capturing the full depth diversity of New Zealand's seismotectonic environment including shallow Alpine Fault ruptures, intermediate-depth Hikurangi subduction events, and deep Wadati–Benioff zone earthquakes.

B. Tiered Quality Distribution

Waveform quality analysis revealed distinct performance tiers across the pipeline stages (Fig. 2). Stage 1, representing the highest-confidence PhaseNet detections (probability $\geq$0.9) cross-validated within $\pm$1 s of TauP predictions, produced 553 waveforms (3.5% of total records). Stage 2 (moderate PhaseNet confidence, 0.5–0.9, $\pm$2 s tolerance) contributed 1,505 waveforms (9.5%). Stage 3, using EQTransformer as primary detector with multi-model consensus, added 701 waveforms (4.4%). Together, Stages 1–3 yielded 2,759 cross-validated waveforms representing 17.4% of all processed records. The remaining 13,084 waveforms (82.6%) failed all automated validation thresholds and are classified as noise-dominated records retained for noise characterization [9], [10].

The high rejection rate (82.6%) reflects the intrinsically challenging noise environment of MEMS-based sensors deployed in urban and suburban settings. Unlike traditional broadband seismometers installed in seismically quiet vaults, Raspberry Shake stations operate in homes, offices, and schools where anthropogenic noise sources traffic, building vibration, HVAC systems, and human activity dominate the frequency bands occupied by local earthquake P- and S-wave arrivals. Retaining these noise records in the dataset is deliberate: they provide realistic negative examples for discriminative training and support noise-augmentation strategies critical for developing robust MEMS-calibrated phase pickers.

C. Manual Validation

Manual validation was performed as a final verification step applied to all waveforms passing at least one automated stage. Each candidate waveform was visually inspected to assess three criteria: (i) overall waveform coherence and absence of clipping or saturation artifacts; (ii) clarity and impulsivity of the P-wave onset; and (iii) temporal consistency of P- and S-wave arrivals relative to event origin time and TauP theoretical predictions (Fig. 3).

Waveforms were excluded during manual review if they exhibited emergent or indistinct phase onsets that precluded confident pick placement, amplitude clipping that distorted the waveform morphology, or arrivals deviating from expected moveout in a manner inconsistent with known seismotectonic structure. Particular attention was paid to distinguishing genuine seismic P-wave onsets from common MEMS artifacts: high-frequency anthropogenic transients (sharp impulsive bursts from nearby machinery or footsteps), tilt noise producing long-period ramps on horizontal MEMS channels,

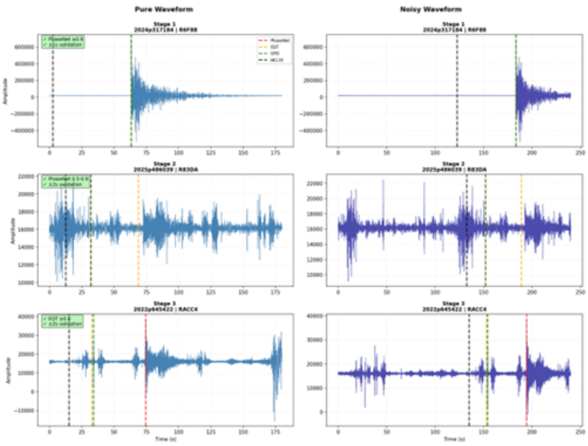

Fig. 3: Comparison of seismic waveforms across processing pipeline stages. Left: pure waveforms (180 s). Right: noisy waveforms (240 s) with pre-event noise. Rows show Stage 1 (PhaseNet $\geq$0.9, $\pm$1 s), Stage 2 (PhaseNet 0.5–0.9, $\pm$2 s), and Stage 3 (EQT $\geq$0.8, $\pm$2 s). Dashed lines: PhaseNet (red), EQTransformer (orange), GPD (green), AK135 (black). Data from Raspberry Shake stations, New Zealand, 100 Hz.

and electronic interference manifesting as narrowband spectral spikes that can superficially resemble impulsive P-wave energy. These artifacts are especially problematic in Stage 2 and Stage 3 records where automated picker confidence is lower, increasing the risk of spurious detections.

Following manual review, 918 waveforms from Stages 1 and 2 were retained as the highest-confidence subset of the dataset, with phase picks verified by both automated consensus and human inspection. These 918 records are directly suitable as training data for machine learning phase pickers. The dataset's geographic diversity spanning the Hikurangi subduction zone, Alpine Fault system, and deep intraslab seismicity (Fig. 1) validates the pipeline's ability to detect and correctly label arrivals across a range of source mechanisms, focal depths, and wave propagation paths [13].

V. Discussion and Conclusion

This work presents a comprehensive MEMS-based earthquake dataset for New Zealand from the Raspberry Shake network, comprising 16,035 four-component waveform records with complete metadata. The multi-tiered validation pipeline yielded 2,759 cross-validated waveforms, of which 918 high-confidence records (Stages 1–2) are directly suitable for training machine learning phase pickers, with remaining records retained for noise characterization. This demonstrates that rigorous quality control applied to citizen-science MEMS recordings can produce training-ready seismic data at a scale not previously available for low-cost instrumentation. The dataset enables development of phase pickers calibrated specifically to MEMS noise characteristics, addressing the known limitation of broadband-trained models generalizing poorly to low-cost sensors. Experience from analogous citizen-science deployments in Bucharest [14] and Haiti further confirms that MEMS stations can yield scientifically credible data under rigorous quality control, lowering the barrier for earthquake monitoring in under-resourced regions [8]–[10].

The structure of the five-stage pipeline was shaped directly by the characteristic noise profiles of Raspberry Shake stations in New Zealand's urban environments, and understanding this relationship is key to interpreting both the tier thresholds and the 82.6% rejection rate. The MEMS channels carry a significantly elevated noise floor compared to broadband instruments below 10 Hz, MEMS self-noise exceeds the low-noise model by 20–40 dB [8] so for small events (M 2.7–3.5), P-wave energy on MEMS channels is often buried in sensor self-noise while the EHZ geophone may still show a clear arrival. The pipeline therefore bases cross-validation on temporal consistency rather than amplitude thresholds, avoiding SNR gates that would disproportionately reject valid low-magnitude records. Three site-specific noise sources further motivated the specific threshold choices at each stage. *Tilt noise* on horizontal MEMS channels from building sway and thermal expansion produces long-period transients that cause pickers to mistime S-wave arrivals; the relaxed $\pm$2 s tolerance in Stages 2–5 (versus $\pm$1 s in Stage 1) accommodates this additional uncertainty. *Anthropogenic broadband transients* footsteps, door slams, and HVAC vibration in the 5–50 Hz band are morphologically similar to P-wave onsets, which drove the requirement for multi-model consensus: PhaseNet, EQTransformer, and GPD were each trained on distinct data with different anthropogenic noise compositions, making coincident false triggers across all three statistically unlikely. The 0.9 probability threshold in Stage 1 excludes most single-transient false detections, while the lower 0.5 threshold in Stage 2 is compensated by mandatory AK135 cross-validation. *Electronic interference* from power supplies and Wi-Fi routers introduces narrowband 50 Hz contamination that can suppress GPD confidence on otherwise clear waveforms; Stage 3's use of EQTransformer as the primary detector exploits its attention mechanism's greater resilience to narrowband spectral artifacts [4]. This spatial noise heterogeneity is preserved in the dataset metadata, enabling future studies to stratify training data by urban noise intensity.

Developing phase pickers robust to high-noise, low-SNR environments requires dedicated noise-augmentation strategies and remains non-trivial. However, success in this domain would be highly impactful: the low cost and ease of deployment of MEMS sensors means that a well-performing picker could scale to thousands of stations globally, far beyond what traditional broadband infrastructure allows. Future work will focus on noise-robust deep learning architectures, network expansion, improved depth estimation, and integration of phase association models to enhance automated event detection. The SeisBench framework provides a useful point of reference for organizing such multi-model, multi-tier validation workflows, and future releases of this dataset are intended to adopt a compatible format to maximize interoperability.

REFERENCES

[1] S. M. Mousavi, Y. Sheng, W. Zhu, and G. C. Beroza, "STanford EArthquake Dataset (STEAD): A global data set of seismic signals for AI," *IEEE Access*, vol. 7, pp. 179464–179476, 2019.
[2] A. Michelini, S. Cianetti, S. Gaviano, C. Giunchi, D. Jozinović, and V. Lauciani, "INSTANCE – the Italian seismic dataset for machine learning," *Earth Syst. Sci. Data*, vol. 13, pp. 5509–5544, 2021.
[3] M. Zhao, Z. Xiao, S. Chen, and L. Fang, "DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology," *Earthq. Sci.*, vol. 36, pp. 84–94, 2023.
[4] S. M. Mousavi, W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza, "Earthquake transformer an attentive deep-learning model for simultaneous earthquake detection and phase picking," *Nat. Commun.*, vol. 11, p. 3952, 2020.
[5] Z. E. Ross, M.-A. Meier, E. Hauksson, and T. H. Heaton, "Generalized seismic phase detection with deep learning," *Bull. Seismol. Soc. Am.*, vol. 108, pp. 2894–2901, 2018.
[6] W. Zhu and G. C. Beroza, "PhaseNet: A deep-neural-network-based seismic arrival-time picking method," *Geophys. J. Int.*, vol. 216, pp. 261–273, 2019.
[7] H. P. Crotwell, T. J. Owens, and J. Ritsema, "The TauP Toolkit: Flexible seismic travel-time and ray-path utilities," *Seismol. Res. Lett.*, vol. 70, pp. 154–160, 1999.
[8] A. D'Alessandro, S. Scudero, and G. Vitale, "A review of the capacitive MEMS for seismology," *Sensors*, vol. 19, p. 3093, 2019.
[9] P. Pierleoni *et al.*, "Low-cost MEMS accelerometers for earthquake early warning systems: A dataset collected during seismic events in central Italy," *Data Brief*, vol. 53, p. 110148, 2024.
[10] V. Cascone, J. Boaga, and G. Cassiani, "Small local earthquake detection using low-cost MEMS accelerometers: examples in northern and central Italy," *Seismic Rec.*, vol. 1, pp. 20–26, 2021.
[11] R. N. Nof *et al.*, "MEMS Accelerometer Mini-Array (MAMA): A low-cost implementation for earthquake early warning enhancement," *Earthq. Spectra*, vol. 35, pp. 21–38, 2019.
[12] W. T. Pike *et al.*, "A vibrating beam MEMS accelerometer for gravity and seismic measurements," *Sci. Rep.*, vol. 10, p. 10415, 2020.
[13] GeoNet, "GeoNet geological hazard information for New Zealand," https://www.geonet.org.nz, 2024.
[14] D. Toma-Danila *et al.*, "Seismic observations in Bucharest area with a Raspberry Shake citizen science network," *Appl. Sci.*, vol. 13, p. 5646, 2023.
[15] M. Beyreuther *et al.*, "ObsPy: A Python toolbox for seismology," *Seismol. Res. Lett.*, vol. 81, pp. 530–533, 2010.
[16] Y. Ni *et al.*, "Curated Pacific Northwest AI-ready seismic dataset," *Seismica*, vol. 2, p. 368, 2023.
[17] Raspberry Shake, "Raspberry Shake [Dataset]," https://doi.org/10.7914/SN/AM, 2016.