Edge seismic phase picking: model distillation for multimodal early-warning systems

Hongyu Xiao[1*], Jacob I. Walter[1], Paul Ogwari[1]

[1]Oklahoma Geological Survey, University of Oklahoma, Norman, OK, USA

Corresponding author: Hongyu Xiao

Email: hongyu.xiao-1@ou.edu

Abstract

Seismic monitoring is entering a high data volume, multimodal sensing regime driven by dense sensor deployments, including Micro-Electro-Mechanical Systems (MEMS) arrays, large-N (100s-1000s) Nodal arrays, Distributed Acoustic Sensing (DAS), and other geophysical instruments. Machine learning models, with appropriate tuning and calibration, are capable of achieving close to expert-level performance in seismic detection and phase picking. The emerging operational challenge is how to scale that machine intelligence across future multimodal sensing networks. Effective incorporation of these capabilities remains constrained by telemetry bandwidth, centralized computation architectures, and communication latency. Here, we present XiaoNet, a proof-of-concept edge machine learning architecture for seismicity monitoring based on a simple system principle: perform first-pass inference near the sensor and transmit concise metadata rather than continuous streams. Using knowledge distillation, we compress a state-of-the-art PhaseNet model by over 93% into XiaoNet, a 240 KB model designed for edge deployment, while preserving strong phase-picking performance across datasets. Deployment on a Raspberry Pi 5 shows sub-millisecond inference latency for 30-second 3-channel waveform windows. By shifting inference to the edge and transmitting concise event metadata payloads, the framework demonstrated the potential for edge inference to reduce communication latency. Substituting an event metadata message payload

for continuous raw waveform stream largely mitigates the dominant source of variable telemetry delay under moderate to heavy load. This result supports a feasible architecture in which edge inference devices complements centralized observatories, potentially improving resilience and response speed for seismic monitoring and may extend to future multimodal subsurface and infrastructure sensing systems.

Significance Statement

High performance machine learning models for geophysical sensing are typically developed and deployed on the cloud or server side, but future monitoring systems will incorporate many more sensors and modalities, from MEMS arrays to DAS and other continuous high-volume data streams. With rapidly increasing data volume, centralizing all raw data could become a bottleneck and point of fragility. We demonstrate that shrinking strong seismic models into edge-deployable models enables local, real-time interpretation with minimal telemetry latency and excels in bandwidth limited use cases. In our case study, this design preserves high detection performance while reducing communication burden and alert latency. On the strength of these results, we envision a hybrid operational setting in which multimodal sensing networks combine edge-based intelligence for rapid first response with centralized systems for deeper analysis, improving speed, resilience, and scalability across geophysical hazard monitoring applications.

Introduction

Earth science is entering an era of distributed sensing at unprecedented scale. Dense networks of accelerometers, Distributed Acoustic Sensing (DAS) systems, nodal arrays, and other continuous high-volume instruments are expanding our ability to monitor dynamic earth processes in real-time, at spatial resolutions previously unattainable with traditional observational networks. In recent years, machine learning has produced highly capable models for earthquake detection, seismic phase picking, association and hazard characterization [1-7]. However, algorithmic capability alone is insufficient to realize operational impact. A critical gap is infrastructure: a powerful model running on a remote server or cloud is of limited use if the raw waveform data required to run it cannot reach that server rapidly and reliably. As sensing networks grow denser [8,9] and more diverse, streaming raw, high-rate data from every channel to centralized servers introduces avoidable latency, network congestion, and a dependence on continuous connectivity, which hampers detection exactly when resilient first response is most critical.

This scaling challenge is particularly visible in induced seismicity settings, where events are often driven by subsurface fluid injection, hydraulic fracturing, geothermal energy extraction and carbon sequestration. These anthropogenic activities have altered seismic hazard in regions such as the U.S. Midcontinent [10-14]. Effective and rapid monitoring of these localized stress perturbations requires dense spatial sampling to track fluid migration and potential fault activation. Traditional broadband stations remain crucial for high-fidelity regional and global analysis, but their sparse geometry can miss fine-scale, high-frequency dynamics. Dense large N Array networks, such as MEMS arrays, offer a practical complement [15-20], and DAS deployments further increase spatial coverage and channel count [21,22]. These advances substantially improve observational capability, but each additional sensor tier compounds the telemetry and real-time processing burden.

Current operations stream continuous multicomponent waveform data from every node to central hubs [23-25]. Recent studies in operational seismology increasingly demonstrate the integration of DAS, machine learning and real-time monitoring infrastructure. Biondi et al. [26] developed a real-time processing framework that integrates DAS streaming and machine-learning based travel-time picking into operational earthquake monitoring infrastructure. Tepp et al. [27,28] demonstrated the potential of DAS observations and machine-learning phase picking to augment regional monitoring. Gou et al. [29] developed an operational DAS-integrated earthquake early warning framework (dEPIC) that combines machine-learning phase picking with event location and magnitude estimation. These developments highlight a broader shift toward integrating dense sensing, machine learning, and real-time processing within operational seismic networks.

At the same time, machine-learning based seismic monitoring is moving beyond independent phase picking toward integrated network-level detection, picking, and phase association. McBrearty and Beroza [30] introduced a graph neural network approach for associating seismic phase arrivals across networks. Sun et al. [31] developed a multi-station phase-picking framework that incorporates spatial and temporal information across the network rather than treating each station independently. These approaches suggest that future seismic monitoring systems will increasingly combine local waveform interpretation with network-level information for event characterization.

At scale, centralized transmission can saturate cellular and satellite links and it can introduce compounding queueing delays. To address this mismatch, inference can be redistributed across the network hierarchy [32,33]. Centralized systems remain essential for raw waveform archiving, relocation, and subsequent geophysical analysis, but first-pass filtering, detection, and phase picking can be performed near the sensor using compact models. This edge-first strategy directly reduces communication volume and opens the door for a future automatic and semi-automatic processing at multiple stages along the telemetry pipeline. Instead of transmitting continuous raw waveform streams, each edge node sends a compact metadata payload for later analysis. Induced seismicity monitoring provides a compelling test case, as shallow earthquakes near energy infrastructure and urban areas require rapid, local response that centralized pipelines may not reliably deliver [24,34].

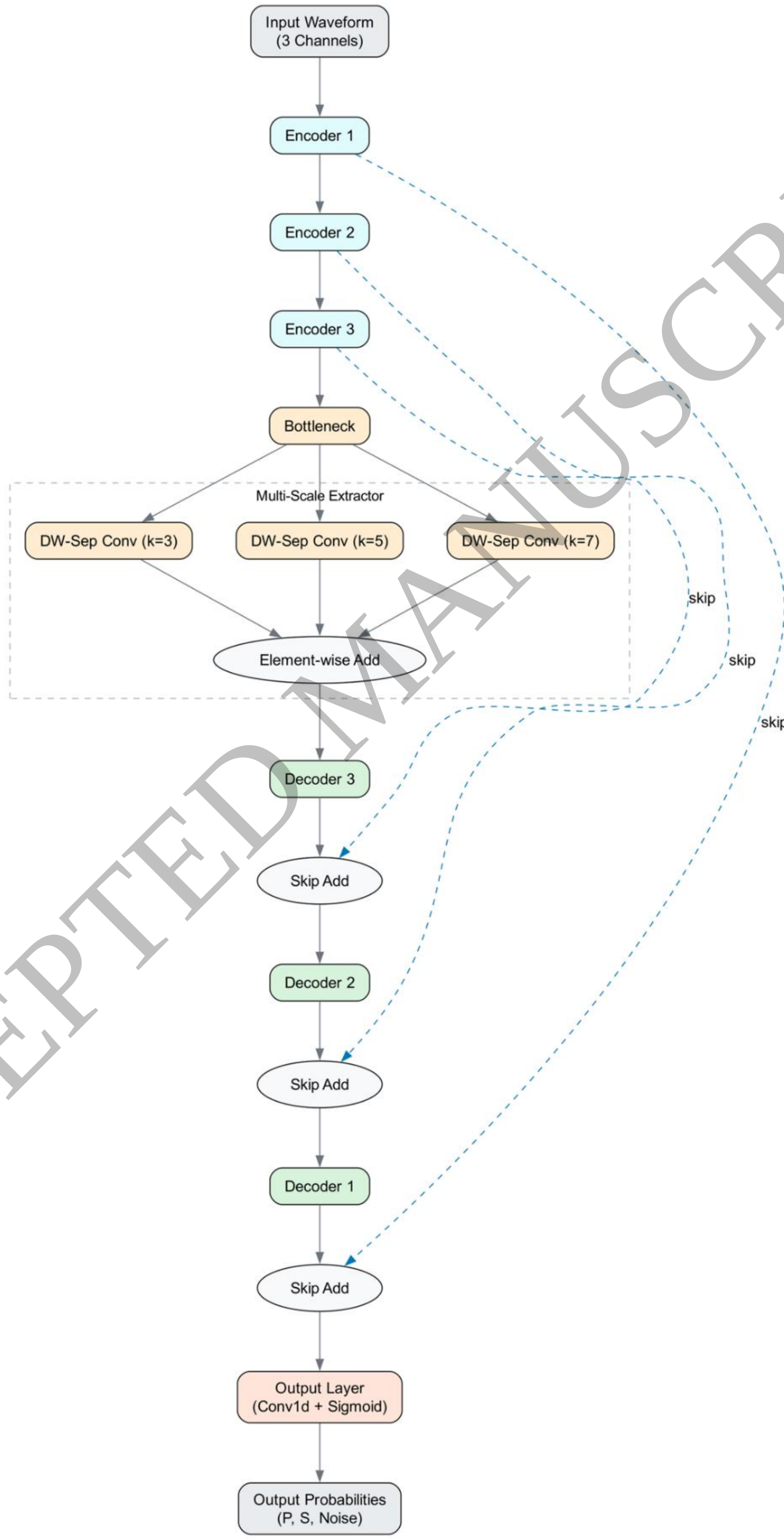

Figure 1. Architecture of the XiaoNet. This model follows a symmetric encoder–decoder topology with three encoder stages performing progressive 4× temporal downsampling via grouped depthwise separable 1D convolutions, each followed by batch normalization and HardSwish activation. A bottleneck layer feeds into a multi-scale feature extractor that fuses parallel convolution paths at kernel sizes 3, 5, and 7 to capture seismic arrivals across multiple temporal scales. The decoder reconstructs temporal resolution with additive skip connections from corresponding encoder stages preserving phase-timing precision. A final 1×1 convolution projects to three output channels (P-arrival, S-arrival, noise), each passed through a sigmoid activation to produce independent per-sample probability estimates.

Knowledge Distillation (KD) provides a model compression framework in which a compact "student" model is trained to imitate the behavior of a larger, high performance "teacher" model [35,26,36]. The central design principle is to construct a student architecture that is sufficiently expressive to approximate the teacher's input-output mapping, while remaining lightweight enough for deployment under strict memory and computation constraints. This is achieved by designing a distillation loss that transfers "knowledge" from the teacher to the student by encouraging agreement between their predictive outputs, rather than relying solely on ground truth labels. Following this design, KD shifts machine learning from direct supervised learning to model behavioral imitation, enabling the student to inherit the functional structure of the teacher model. The resulting model is significantly smaller while preserving much of the teacher's event relevant performance, making inference feasible in resource-constrained environments such as edge computing systems.

With architectural innovations such as depthwise separable convolutions [37], which substantially reduce the number of parameters and computational operations relative to conventional convolutions, we build a compact, deployable seismic model (**Figure 1**). We first evaluate the performance of this new XiaoNet model architecture on widely used benchmark earthquake datasets. We then use the pretrained PhaseNet model to improve XiaoNet through the aforementioned KD framework. The final distilled student model is approximately 240 KB in size while preserving strong phase picking performance. We validate real-time phase picking on both a Raspberry Pi 5 and Jetson Orin Nano edge node and demonstrated the potential for edge inference to reduce communication latency by performing first-pass inference locally rather than continuous waveform streaming. The distilled model was also deployed on an ESP32-S3 edge computing device to verify lightweight portability, although without detailed latency benchmarking. Together, these results provide a foundation for distributed edge inference across heterogeneous sensing platforms. In a future hybrid architecture, MEMS arrays, DAS deployments, and additional sensors could operate on appropriate hardware platforms, each contributing structured local machine intelligence. This could enable distributed first-response and centralized high-fidelity analysis to operate as complementary layers of a unified architecture.

Results

Efficacy of Knowledge Distillation on Diverse Datasets

We first established the baseline performance and generalization of XiaoNet by training the compact model from scratch, without pretrained weights or knowledge distillation. We trained it independently across diverse tectonic and induced seismicity datasets: STanford EArthquake Dataset (STEAD) [38], Swiss Seismological Service (SED) dataset (ETHZ) [39], the Texas Seismological Network (TXED) [9], and the Oklahoma Labeled AI Dataset (OKLAD) [40] derived from the Oklahoma Geological Survey (OGS) network [8]. XiaoNet trained from scratch demonstrated strong phase-picking performance across these datasets. It displayed high phase picking capability on the global STEAD dataset (**Figure 2A**), with a True Positive Rate [TPR] of 95.58% for P-arrivals and 94.67% for S-arrivals. It also demonstrated strong P-arrival picking capability on ETHZ (91.38% TPR). However, performance was comparatively constrained on complex induced seismicity datasets, yielding reasonable but lower metrics for TXED (80.16% P TPR, 82.99% S TPR) and OKLAD (81.23% P TPR, 82.41% S TPR). TXED and OKLAD are predominantly composed of region-specific induced seismicity associated with subsurface fluid injection, hydraulic fracturing, and related anthropogenic activities, making them challenging test cases characterized by heterogeneity, strong local scattering, and elevated anthropogenic noise. These results demonstrate that the compact XiaoNet architecture can independently learn effective P- and S-phase representations without knowledge transfer, while also highlighting the benefit of subsequent knowledge distillation for recovering performance under challenging regional conditions.

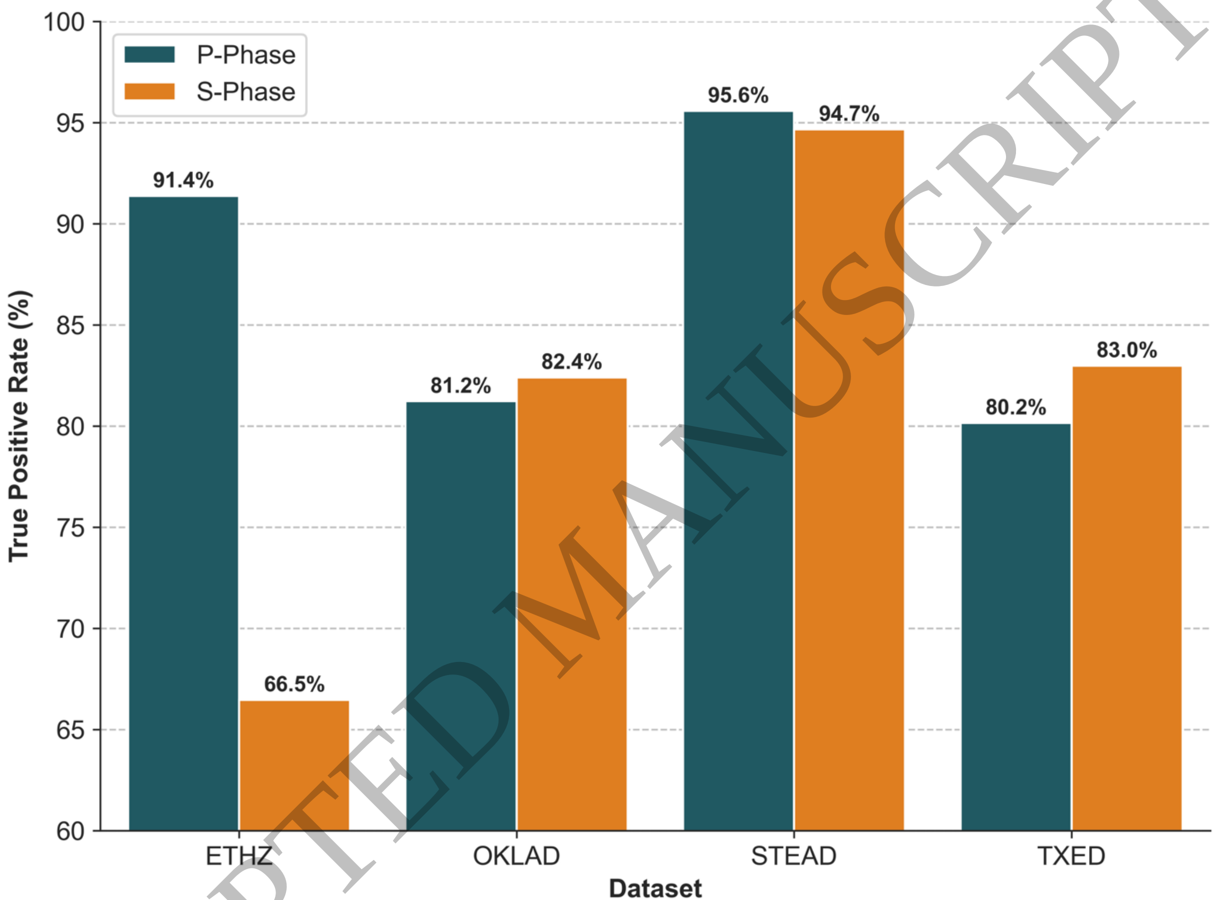

Figure 2A. True positive rates for P-arrival and S-arrival detection when XiaoNet is trained from scratch across four seismological datasets. XiaoNet achieves high phase picking performance on the global tectonic STEAD dataset (P TPR: 95.58%, S TPR: 94.67%) and strong P-arrival performance on ETHZ (P TPR: 91.38%), but yields reasonably lower performance on the induced seismicity datasets TXED (P TPR: 80.16%, S TPR: 82.99%) and OKLAD (P TPR: 81.23%, S TPR: 82.41%). The performance disparity between tectonic and induced seismicity regimes highlights the difficulty of extracting body-wave arrivals from shallow, anthropogenic dominant events.

To improve XiaoNet performance on challenging induced seismicity settings, we implemented a knowledge distillation (KD) framework using pre-trained PhaseNet as the teacher model. We first trained the PhaseNet teacher from scratch using the full OKLAD dataset, achieving P-wave and S-wave TPRs of

88.1% and 87.0%, respectively. We then distilled this knowledge into the same XiaoNet architecture, using XiaoNet as the student model and varying the proportion of OKLAD training data from 0.5% to 100% to evaluate how performance scales with the amount of distillation data.

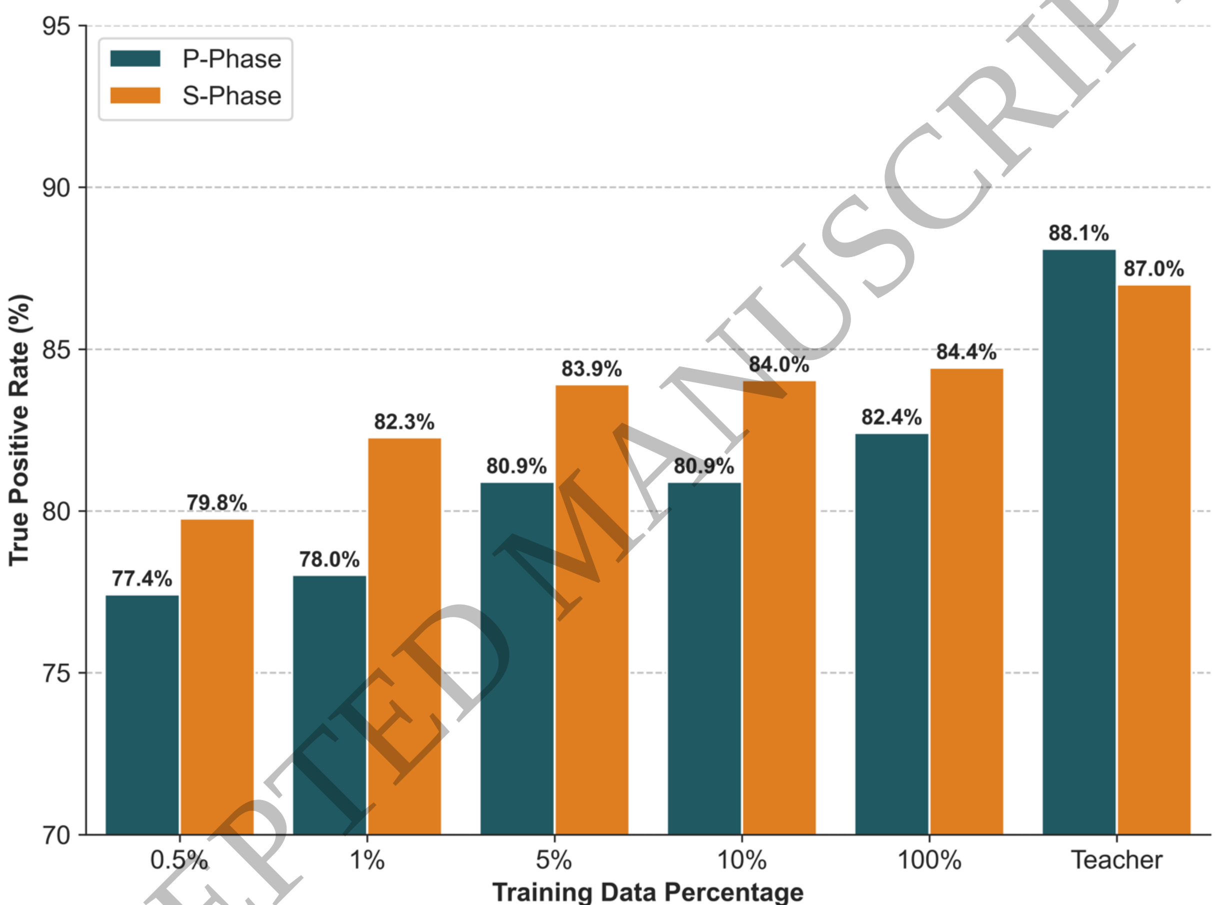

Figure 2B. Progressive improvement in P arrival and S arrival true positive rates as a function of OKLAD data fraction used for Knowledge Distillation. The KD Student model (XiaoNet) shows progressive gains in both P and S TPR as the distillation training set scales from 0.5% to 100% of OKLAD (P TPR: 77.42% to 82.41%; S TPR: 79.77% to 84.43%). Notably, even at 0.5%, the student model achieves reasonable detection rates, indicating that the Teacher's structural representations transfer efficiently with minimal labeled examples.

The KD student model, which learns to mimic both the final probabilities and intermediate feature representations of the teacher, successfully recovered strong phase picking performance (**Figure 2B**). The results demonstrated a progressive and scalable improvement in the student model's TPR for both P and S phases. Using only 1% of the OKLAD training data for distillation, the student model achieved TPRs of 78.0% for P-arrivals and 82.3% for S-arrivals, compared with 72.0% and 79.2%, respectively, for XiaoNet trained from scratch using 1% of the data. Scaling to 100% data distillation further advanced the student model's performance to F1-scores of 0.8629 (P-arrival) and 0.8610 (S-arrival) alongside TPRs of 82.41% and 84.43%. This process preserved sensitivity to emergent S-wave arrivals, which are historically more difficult to differentiate from preceding P-wave coda but remain the primary source for catastrophic damage in cities; the rapid characterization of body wave type can lead to improvement in the performance of some earthquake early warning algorithms [32].

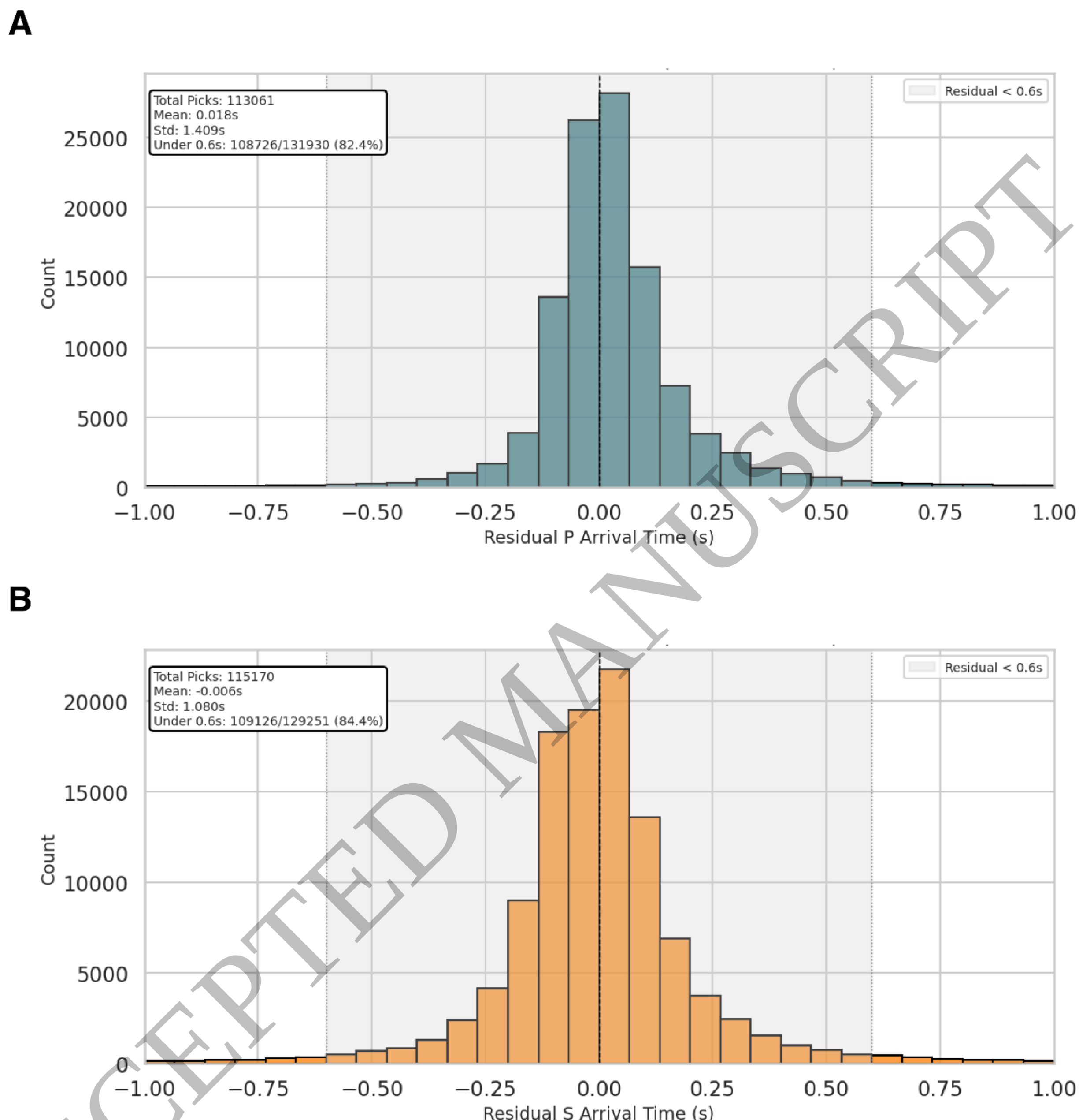

Figure 3. Histograms of phase-picking residuals (−1 to 1 s) between the distilled student model predictions and analyst-labeled ground truth. (A) P-phase picks and (B) S-phase picks.

To further assess phase picking quality and timing accuracy, we evaluated the residuals between model predictions and analyst-labeled P and S phase arrivals. For the Teacher model, 115,826 of 131,483 P arrival picks and 112,648 of 129,503 S arrival picks had residuals within 0.6 s to the ground truth, corresponding to TPRs of 88.1% and 87.0%, respectively. The Teacher predictions exhibited mean residuals of 0.0046 s for P phases and 0.0044 s for S phases, with standard deviations of 1.1244 and 0.7018 s, respectively. The distilled Student model recovered 82.41% of P arrivals and 84.43% of S arrivals within the 0.6 s. The student model's mean residuals were 0.018 s for P phases and −0.006 s for S phases, with standard deviations of 1.409 and 1.080 s, respectively. The near-zero mean residuals indicate minimal systematic timing bias in the student model, while the larger standard deviations relative to the Teacher indicate greater dispersion in the predicted arrival times. The student model residual distributions are shown in **Figure 3**. And waveform and picking examples of Teacher model and Student model predictions are shown in **Figure 4**. These results demonstrate that the distilled Student model preserves near-zero systematic timing bias to both P- and S-phase arrivals, while exhibiting some increased variability in individual pick times.

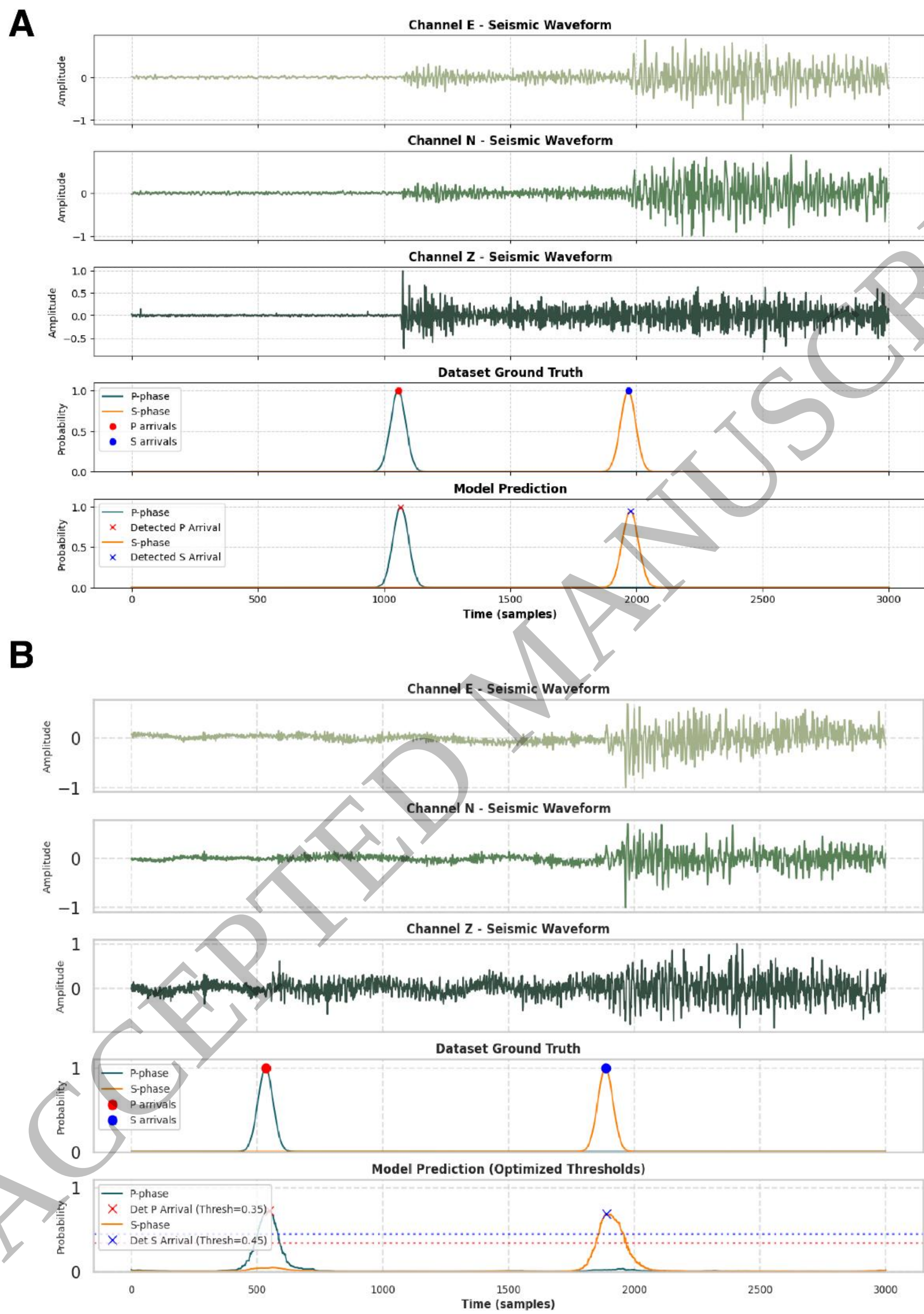

Figure 4. Examples of waveform data, ground truth label, and model predictions. (A) Three-component waveforms (E, N, and Z), OKLAD analyst-labeled P and S phase arrivals, and Teacher model predictions. (B) Example of student model predictions. Red and blue markers denote P- and S-phase arrivals, respectively; dots indicate ground-truth labels and crosses indicate model predictions.

Extreme Hardware Compression and Hybrid Capability

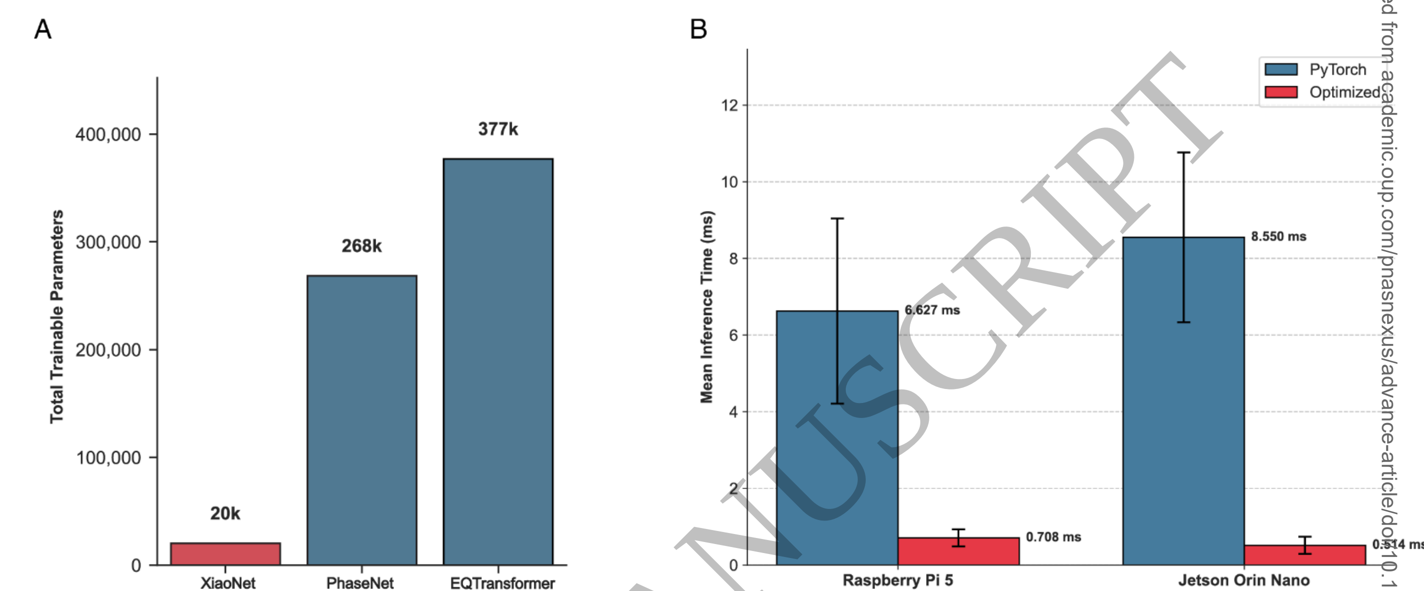

Figure 5A. Comparison of Total Trainable Parameters between XiaoNet, PhaseNet and EQtranformer. Figure 5B. Single trace inference runtime of XiaoNet across two edge hardware platforms (Pi 5 and Jetson Orin Nano) using standard and optimized execution formats.

The KD student model (XiaoNet) contains 20,208 trainable parameters and achieves high accuracy with a compact architecture. Compared to PhaseNet (268,443 total trainable parameters) and EQTransformer (376,935 total trainable parameters), this represents a 92.4% and 94.6% reduction in model size, respectively (**Figure 5A**). The resulting memory footprint is approximately 240 KB, allowing the full model to fit within the L1/L2 cache of modern edge processors. This reduces dependence on external memory access during inference and mitigates associated latency overheads.

To validate real-time stream processing on hardware that can reasonably be integrated at the edge, the model was benchmarked on accessible edge computing hubs (Raspberry Pi 5 and Jetson Orin Nano) using both standard PyTorch (`.pth`) and optimized formats (ONNX and TensorRT accordingly). While standard execution yielded mean latencies of 6.63 ms (Pi 5) and 8.55 ms (Orin Nano), hardware-specific optimization reduced mean inference times to 0.71 ms and 0.51 ms, respectively, for a 30-second (100 Hz) 3-channel waveform window (**Figure 5B**). These sub-millisecond latencies confirm that high efficiency edge hubs can act as local aggregators for multiplexed sensor clusters without bottlenecking the alert sequence.

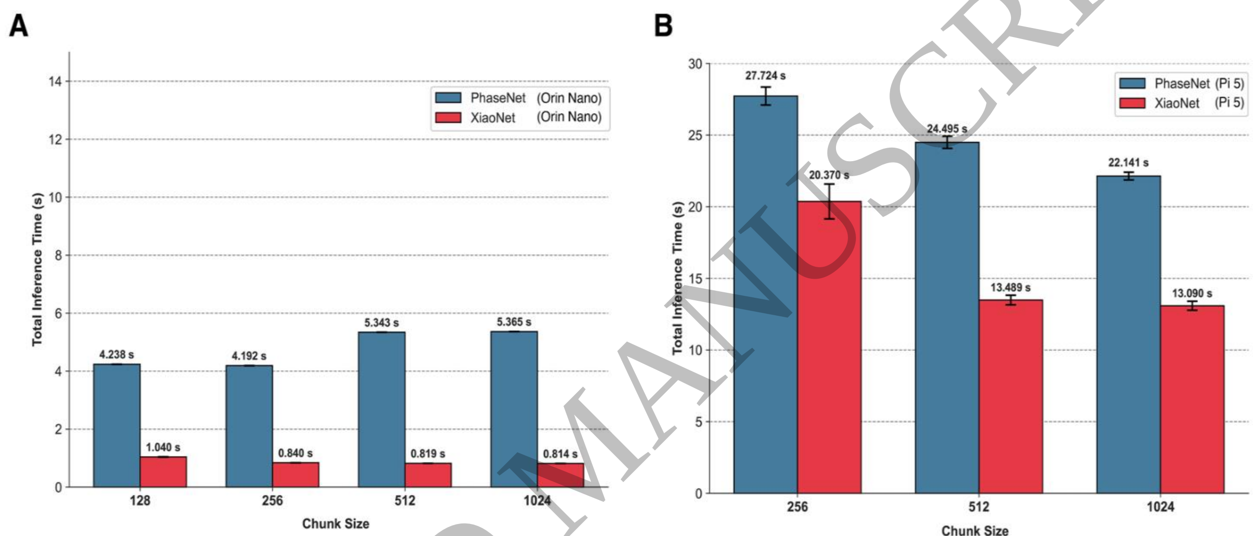

Figure 6. Benchmarked real-time edge streaming inference time across a 10,000-channel DAS array. (A) NVIDIA Jetson Orin Nano and (B) Raspberry Pi 5. For each edge computing platform, total inference time was measured for a continuous 30-second time window (3,001 samples at 100 Hz across 10,000 channels with a 15-s hop, corresponding to 50% overlap), comparing XiaoNet with standard PhaseNet over 20 consecutive streaming runs. Bars and error bars represent the mean and standard deviation, respectively, across varying processing batch chunk sizes.

Furthermore, to evaluate the model's potential for scalable inference on ultra-dense networks such as DAS, we benchmarked continuous streaming over a 10,000-channel DAS array using 100-Hz sampling rate, 30 s processing trace length and 15 s sliding hops (50% overlap). Each channel was augmented with two additional dimensions to match the model input format. On the NVIDIA Jetson Orin Nano (**Figure 6A**), XiaoNet (PyTorch in pth format) processed the full 10,000-channel input in 0.814 s, achieving a 6.6 times speedup over PhaseNet (5.365 s) and consuming only 5.4% of the 15.0 s hop budget (the maximum processing time available before the next input window arrived), leaving more than 94.5% real-time headroom. On the Raspberry Pi 5 **(Figure 6B)**, XiaoNet (ONNX format) completed the full 10,000-channels input inference in 13.09 s, satisfying the 15.0 s streaming threshold, whereas PhaseNet required 22.14 s–27.72 s, exceeding real-time budget. These results demonstrated the feasibility of real-time, array-wide edge seismic monitoring using lightweight architectures on accessible commercial edge computing devices. We note that this benchmark evaluates computational efficiency and does not constitute a direct comparison with DAS specific phase-picking models such as PhaseNet-DAS [39]. Therefore, the results are intended to demonstrate the edge-inference scalability of XiaoNet rather than to establish its superiority over models specifically optimized for DAS data.

Most significantly, such tiny footprint provides an implementation pathway for integrating directly onto bare-metal MEMS microcontrollers. We successfully deployed XiaoNet on an ESP32-S3 microcontroller, proving the model is loadable and stable even when handling long-duration 30-second (100 Hz) multi-component traces. Hardware testing demonstrated that the ESP32-S3 can process a 30-second window in 2,148 ms, representing a processing speed faster than real-time. This demonstrates that modern seismic phase picking capability is no longer restricted to cloud supercomputers, which can enable thousands of ultralow power micro-processors to seamlessly augment the sparse network of traditional back bone stations.

Spatial and Temporal Impact on Early Warning

State-of-health analysis of the current Oklahoma Geological Survey monitoring network shows a median telemetry latency of 1.8 seconds embedded within the waveform delivery pipeline.[8] This delay arises because traditional cloud centric architectures must transmit full raw waveform payloads (~27 KB per 30-second, 100 Hz, 24-bit window) over cellular links (LTE/4G) before centralized servers can begin phase picking. The packetization of data and acknowledgement of receipt required to send over Transmission

Control Protocol (TCP) can further increase telemetry delays. Every delay within the data transmission pipeline potentially hinders earthquake early warning system efficacy.

XiaoNet-based models eliminate this bottleneck by performing inference directly at the sensor edge. Rather than streaming raw waveforms, the edge node completes phase picking locally and transmits only a compact 32-byte metadata payload (pick timestamp, wave type, probability, and station ID) via lightweight MQTT protocols [41] —a payload reduction exceeding 99.8%. This compression effectively mitigates transmission congestion and entirely removes centralized queuing from the alert pipeline.

The practical implication is that the reduced processing time can effectively translate into additional warning time under idealized propagation assumptions. For typical crustal S-wave velocities (~3.5 km/s), a reduction of 1.8 seconds corresponds to approximately 6 km of additional wave travel distance prior to the arrival of ground motion. In induced seismicity settings, where hypocenters are often shallow and critical infrastructure may lie within ~10 km of the source, even modest reductions in processing time can increase the available window for alert dissemination before the onset of strong shaking. Deploying low-latency edge sensors alongside existing broadband networks therefore has the potential to extend the effective operational reach and performance of real-time monitoring systems.

Discussion

The central finding of this work is that Knowledge Distillation can compress a research level seismic phase picker into a model small enough to execute on edge computing devices while retaining detection performance under the evaluated conditions. This result has direct implications for the design and operation of dense seismic networks (Large-N arrays): edge ML models could support a hybrid architecture in which distributed sensors provide rapid, autonomous first-pass detections that supplement the high fidelity, high dynamic range backbone network analysis performed by centralized systems.

The Knowledge Distillation results demonstrated an instructive asymmetry. XiaoNet trained independently achieves high fidelity on the global STEAD dataset but requires Teacher guidance to recover comparable performance on regional induced seismicity data (OKLAD), suggesting that the compact architecture readily captures broadband waveform features but benefits from supervised transfer for the subtler arrivals' characteristic of shallow, noisier local induced events. This asymmetry carries a practical implication: the distillation pipeline is portable. Operators seeking to deploy edge inference models in a new tectonic setting need only retrain the Teacher on relatively small amount of locally labeled data. And the student architecture and distillation protocol remain unchanged.

Importantly, the progressively smaller gains in TPR performance with increasing training data volume **(Figure 2B)** suggest diminishing marginal returns from additional training data. With smaller training data volumes, each additional subset of the OKLAD dataset provides substantial new information, enabling the student model to better approximate the Teacher's decision and representations. However, because the Teacher already provides a strong supervisory performance, a relatively small amount of localized data may be sufficient for the student to learn the characteristics of the target region/dataset phase-picking ability. As the training volume increases, newly added training samples increasingly contain patterns that are already represented in the previous training data, providing less new information for further performance refinement. Consequently, increasing the training data volume continues to improve phase-picking performance, but the degree of improvement decreases as the student approaches the performance attainable from the available data and distillation architecture. This indicates that substantial knowledge transfer can be achieved with a relatively small fraction of the localized data, while larger training volumes can provide progressively finer improvements in sensitivity and generalization.

Importantly, this data efficiency further supports the portability of the proposed distillation framework, as adapting the student model to a new tectonic setting may not require extensive localized data.

The hardware benchmarks demonstrated deployment feasibility across distinct operational tiers. Sub-millisecond optimized inference on the Raspberry Pi 5 and Jetson Orin Nano shows that single board computers can effectively serve as local aggregators for multiplexed sensor clusters without contributing noticeable latency to the alert pipeline. The successful execution on an ESP32-S3 device, which processed a full 30 second, 100 Hz trace faster than real-time, is more noteworthy, because it further demonstrates the deployment on the class of ultralow power processors already integrated into MEMS accelerometers. Phase picking at the point of measurement, rather than after telemetry, represents a qualitative shift in network topology.

The architecture also has potential applicability beyond MEMS based sensing. Distributed Acoustic Sensing (DAS) systems can generate high-dimensional, continuous waveform streams, making localized inference particularly attractive for reducing the volume of data that must be transmitted upstream. Our 10,000-channel edge streaming benchmark (**Figure 6**) demonstrates the computational feasibility of applying the lightweight XiaoNet architecture to such high channel-count data streams on commercially accessible edge computing devices. However, this benchmark evaluates computational scalability rather than phase-picking accuracy on DAS recordings and does not constitute a comparison with DAS-specific phase-picking models. DAS-specific model development and validation represent a prospective extension of this framework. More broadly, the principle of performing inference close to where data are generated and transmitting only compact phase-picking metadata upstream is potentially applicable across sensing modalities and becomes increasingly valuable as data density increases.

Several limitations are still worth consideration. Domain shift remains a key challenge: models trained across diverse tectonic datasets can degrade performance in other region such as basins or noisy environments that were not represented in the training distribution [43,44]. Federated protocols [45], in which local models adapt to site specific conditions while preserving communication efficiency and data governance, offer a promising path forward. In this setting, each edge deployed model can be further fine-tuned on site-specific seismic characteristics, while a central aggregation process improves global performance across regions without requiring direct access to raw data. Extension to DAS-specific edge models [46], cross-modal event association, and direct integration with operational platforms (i.e. SeisComP, AQMS, and ShakeAlert) represent important next steps.

This study establishes that edge-first seismic phase picking inference is technically viable, operationally promising for an active monitoring network. By reducing telemetry-related latency by at least 1.8 s relative to the cloud-centric pipeline, edge inference provides additional time for downstream event characterization and alert generation. As a first-order estimate, this 1.8-s latency reduction corresponds to approximately 6.3 km of propagation distance, illustrating the potential significance of even modest latency reductions for communities located near induced seismicity sources. For such communities, where available warning time may be particularly limited, reducing processing and telemetry latency could therefore contribute to more effective earthquake early warning.

Materials and Methods

Dataset Configuration

We evaluated the XiaoNet framework across four seismological datasets representing two distinct seismogenic regimes. STEAD [38] and ETHZ [39] served as global tectonic benchmarks, including primarily natural earthquakes across various epicentral distances, focal depths, and magnitude ranges. The STanford EArthquake Dataset (STEAD) is a comprehensive, globally curated benchmark comprising approximately 1.2 million high-quality, three-component seismic waveforms from over 450,000 local earthquakes, specifically designed for large-scale training of machine learning models. And the ETHZ dataset is a manually compiled, regional benchmark containing more than 20,000 high-resolution seismic arrivals recorded across Switzerland and neighboring alpine regions by the Swiss Seismological Service (SED) and AlpArray networks. Meanwhile, TXED [9] and OKLAD [40] provided induced/anthropogenic dominated seismicity benchmarks characteristic of shallow, injection/hydraulic fracking related events. The Texas Earthquake Dataset (TXED) is a specialized regional benchmark consisting of over 20,000 manually labeled earthquake events and 200,000 noise waveforms from the Texas Seismological Network (TexNet). Oklahoma Labeled AI Dataset (OKLAD) comprises over 1 million labeled three-component (3C) waveform traces and was selected as the primary dataset for the Knowledge Distillation experiments, serving as the training data for both the Teacher model and the distillation data efficiency scaling analysis.

All seismic data, including the STEAD, ETHZ, TXED and OKLAD benchmarks, were accessed and processed through SeisBench, an open-source Python framework that provides a unified API for standardizing seismic datasets and deploying deep learning models. [47] All waveforms were resampled to 100 Hz and cut into 30-second windows, providing sufficient duration to capture both P- and S-wave arrivals within a single inference frame. Data were divided into 70% training, 15% validation, and 15% testing splits. Each channel was independently normalized to account for varying instrument calibrations and local site amplification effects.

Methods

All models were trained from random initialization without pretrained weights. PhaseNet [2] was used as the teacher model in a knowledge distillation framework. PhaseNet is 1-dimensional U-Net architecture widely adopted in seismology for its expansive temporal receptive field. It implements a multi-tiered encoder–decoder topology with residual skip connections and symmetric upsampling. This model It accepts 3 channel waveform input (3001 samples) and outputs per sample probability distributions for P-arrival, S-arrival, and noise. PhaseNet was trained from scratch on OKLAD and its model weight was subsequently frozen during student training.

XiaoNet, the student model, is a lightweight encoder–decoder network designed for edge deployment. (Figure 1) The architecture employs grouped depthwise separable 1D convolutions [37] throughout, minimizing parameter count while preserving representational capacity. The encoder consists of three stages with progressive 4× temporal downsampling (kernel sizes 7, 5, 5), each followed by batch normalization and HardSwish activation function [48]. A bottleneck layer with 4× downsampling feeds into a multi-scale feature extractor that fuses parallel convolution paths at kernel sizes 3, 5, and 7 to capture seismic arrivals across multiple temporal scales. The symmetric decoder upsamples via nearest-neighbor interpolation followed by depthwise separable convolutions, with additive skip connections from the corresponding encoder stages to preserve phase-timing precision. The output layer is a 1×1 convolution projecting to three channels (P, S, noise), each passed through a sigmoid activation to produce independent per-class probabilities.

XiaoNet was evaluated both as a baseline model trained independently from scratch and as the student model within the distillation framework. In the baseline setting, XiaoNet was trained using standard supervised learning on each dataset without access to teacher outputs. In the distillation setting, the student model was trained using a composite loss combining ground-truth supervision and teacher guidance.

Specifically, the student model was trained using a two-term composite loss:

$$L_{\text{Total}} = \alpha\, L_{\text{Hard}} + (1-\alpha)\, L_{\text{Soft}} \qquad \text{(e.q.1)}$$

where the Hard Target Loss (L_{Hard}) is binary cross-entropy (BCE) between the student model's sigmoid outputs and the ground-truth analyst labels. And the Soft Target Loss (L_{Soft}) is BCE between the student model's outputs and the Teacher's probability predictions, treated as independent soft labels. When temperature softening is applied (T > 1), the Teacher's output probabilities are first converted back to logits, scaled by 1/T, and passed through a sigmoid function to produce softened targets. The weighting parameter alpha was 0.5 which represents equal influence from ground-truth labels and Teacher guidance. To evaluate data efficiency, the distillation was repeated using 0.5%, 1%, 5%, 10%, and 100% of the OKLAD training partition, with all other hyperparameters held constant.

Hardware validation was performed across three deployment tiers representing constrained edge environments. The distilled model is first evaluated on a Raspberry Pi 5 (quad-core ARM Cortex-A76) using both standard PyTorch inference and ONNX-optimized execution on 30-second, three-component waveform windows. The same inputs are then deployed on an NVIDIA Jetson Orin Nano (128-core GPU, 8 GB LPDDR5), using both PyTorch and TensorRT-optimized inference backends. To assess extreme low-power feasibility, the model is additionally deployed on an ESP32-S3 microcontroller (N16R8).

Execution is subsequently validated under bare-metal conditions to confirm compatibility with ultra-low power MEMS hardware.

To quantify the impact of edge inference on earthquake early warning (EEW) latency,we used the empirically observed median telemetry latency of 1.8 seconds from the Oklahoma monitoring network's state-of-health diagnostics. The traditional cloud pipeline was modeled with a 30-second data integration window followed by raw waveform transmission (~27 KB per window) over LTE/4G, subject to the observed 1.8-second combined transmission and queuing delay, plus server-side inference (~100 ms). The edge pipeline was modeled with identical data integration followed by on-device inference (sub-millisecond on Pi 5) and transmission of a 32-byte MQTT metadata payload (pick timestamp, wave type, confidence, station ID), with negligible transmission latency (~50 ms). The latency difference was converted to an effective spatial warning radius under idealized propagation conditions using an assumed typical crustal S-wave velocity of 3.5 km/s.

Data and Software Availability Resources

Waveform data and associated metadata used in this study were obtained from the Oklahoma Geological Survey (OGS) data archive [3]. The XiaoNet architecture, trained models and distilled models and associated source code used in this study are publicly available on Zenodo: https://doi.org/10.5281/zenodo.22097958. OKLAD curated waveform data are available openly available on Zenodo: https://doi.org/10.5281/zenodo.18991761

Acknowledgments

This research was supported by the Oklahoma Geological Survey (OGS). We also would like to thank the two anonymous reviewers for their thoughtful and constructive comments, which helped improve the manuscript.

Competing Interest

The authors acknowledge that there are no conflicts of interest recorded.

Funding

This work was supported by the Data Institute for Societal Challenges (DISC) through a DISC Postdoctoral Fellowship.

Author Contributions

Conceptualization: HX; Methodology: HX; Formal Analysis: HX, JW; Investigation: HX, JW, PO; Data Curation: HX, JW, PO; Writing – Original Draft: HX; Writing – Review & Editing: HX, JW, PO; Visualization: HX; Supervision: JW; Project Administration: HX, JW; Funding Acquisition: HX, JW.

All authors reviewed and approved the final manuscript.

References

1. Ross, Z. E., Meier, M. A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. Bulletin of the Seismological Society of America, 108(5A), 2894-2901.
2. Zhu, W., & Beroza, G. C. (2019). PhaseNet: a deep-neural-network-based seismic arrival-time picking method. Geophysical Journal International, 216(1), 261-273.
3. Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nature communications, 11(1), 3952.
4. Saad, O. M., Huang, G., Chen, Y., Savvaidis, A., Fomel, S., Pham, N., & Chen, Y. (2021). Scalodeep: A highly generalized deep learning framework for real-time earthquake detection. Journal of Geophysical Research: Solid Earth, 126(4), e2020JB021473.
5. Walter, J. I., Ogwari, P., Thiel, A., Ferrer, F., & Woelfel, I. (2021). easyQuake: Putting machine learning to work for your regional seismic network or local earthquake study. Seismological Society of America, 92(1), 555-563.
6. Hsu, T. Y., & Huang, C. W. (2021). Onsite early prediction of PGA using CNN with multi-scale and multi-domain P-waves as input. Frontiers in Earth Science, 9, 626908.
7. Chen, Y., Saad, O. M., Savvaidis, A., Chen, Y., & Fomel, S. (2022). 3D microseismic monitoring using machine learning. Journal of Geophysical Research: Solid Earth, 127(3), e2021JB023842.
8. Walter, J. I., Ogwari, P., Thiel, A., Ferrer, F., Woelfel, I., Chang, J. C., ... & Holland, A. A. (2020). The Oklahoma geological survey statewide seismic network. Seismological Research Letters, 91(2A), 611-621.
9. Chen, Y., Savvaidis, A., Saad, O. M., Dino Huang, G. C., Siervo, D., O'Sullivan, V., ... & Grigoratos, I. (2024). TXED: The Texas earthquake dataset for AI. Seismological Research Letters, 95(3), 2013-2022.
10. Ellsworth, W. L. (2013). Injection-induced earthquakes. Science, 341(6142), 1225942.
11. Walter, J. I., Frohlich, C., & Borgfeldt, T. (2018). Natural and induced seismicity in the Texas and Oklahoma Panhandles. *Seismological Research Letters*, *89*(6), 2437-2446.
12. Kim, K. H., Ree, J. H., Kim, Y., Kim, S., Kang, S. Y., & Seo, W. (2018). Assessing whether the 2017 M w 5.4 Pohang earthquake in South Korea was an induced event. science, 360(6392), 1007-1009.
13. Ogwari, P., Walter, J. I., Chen, X., Woelfel, I., Thiel, A., Ferrer, F., & Brooks, C. (2025). Hydraulic Fracturing Enhances Wastewater Disposal Seismicity in Eastern Oklahoma. Seismological Research Letters.
14. Murray, K. E., Brooks, C., Walter, J. I., & Ogwari, P. O. (2023). Oklahoma's coordinated response to more than a decade of elevated seismicity.
15. Cochran, E. S., Lawrence, J. F., Christensen, C., & Jakka, R. S. (2009). The quake-catcher network: Citizen science expanding seismic horizons. Seismological Research Letters, 80(1), 26-30.
16. Kong, Q., Allen, R. M., Schreier, L., & Kwon, Y. W. (2016). MyShake: A smartphone seismic network for earthquake early warning and beyond. Science advances, 2(2), e1501055.
17. Bossu, R., Roussel, F., Fallou, L., Landès, M., Steed, R., Mazet-Roux, G., ... & Petersen, L. (2018). LastQuake: From rapid information to global seismic risk reduction. International journal of disaster risk reduction, 28, 32-42.
18. D'Alessandro, A., Scudero, S., & Vitale, G. (2019). A review of the capacitive MEMS for seismology. Sensors, 19(14), 3093.
19. Anthony, R. E., Ringler, A. T., Wilson, D. C., & Wolin, E. (2019). Do low-cost seismographs perform well enough for your network? An overview of laboratory tests and field observations of the OSOP Raspberry Shake 4D. Seismological Research Letters, 90(1), 219-228.
20. Won, J., Park, J., Park, J. W., & Kim, I. H. (2020). BLESeis: Low-cost IOT sensor for smart earthquake detection and notification. Sensors, 20(10), 2963.
21. Ajo-Franklin, J. B., Dou, S., Lindsey, N. J., Monga, I., Tracy, C., Robertson, M., ... & Li, X. (2019). Distributed acoustic sensing using dark fiber for near-surface characterization and broadband seismic event detection. Scientific reports, 9(1), 1328.

22. Zhan, Z. (2020). Distributed acoustic sensing turns fiber-optic cables into sensitive seismic antennas. Seismological Research Letters, 91(1), 1-15.
23. Kuyuk, H. S., & Allen, R. M. (2013). Optimal seismic network density for earthquake early warning: A case study from California. *Seismological Research Letters*, *84*(6), 946-954.
24. Allen, R. M., & Melgar, D. (2019). Earthquake early warning: Advances, scientific challenges, and societal needs. Annual Review of Earth and Planetary Sciences, 47(1), 361-388.
25. Stubailo, I., Alvarez, M., Biasi, G., Bhadha, R., & Hauksson, E. (2021). Latency of waveform data delivery from the southern California seismic network during the 2019 Ridgecrest earthquake sequence and its effect on ShakeAlert. Seismological Society of America, 92(1), 170-186.
26. Ba, L. J., & Caruana, R. (2014). Do deep nets really need to be deep?. Advances in neural information processing systems, 27.
27. Tepp, G., Artiaga, N., & Biondi, E. (2026). Evaluation of distributed acoustic sensing phase pick quality and performance for operational earthquake monitoring. *Seismological Research Letters*.
28. Tepp, G., Yu, E., Bhaskaran, A., Tam, R., Zhu, W., Newman, Z., ... & Scheckel, N. (2025). Improvements from incorporating machine learning algorithms into near real-time operational post-processing. *Scientific Reports*, *15*(1), 28938.
29. Gou, Y., Nof, R. N., Pardini, B., & Allen, R. M. (2025). Integrating fiber-optic seismic arrays into earthquake early warning systems with the dEPIC framework. *Scientific Reports*, *16*(1), 960.
30. McBrearty, I. W., & Beroza, G. C. (2023). Earthquake phase association with graph neural networks. *Bulletin of the Seismological Society of America*, *113*(2), 524-547.
31. Sun, H., Ross, Z. E., Zhu, W., & Azizzadenesheli, K. (2023). Phase neural operator for multi-station picking of seismic arrivals. *Geophysical Research Letters*, *50*(24), e2023GL106434.
32. Zhao, M., Zhuowei, X., & Chen, S. (2022, December). DiTing-IoT: deep-learning-enabled real-time seismic data processing via edge and cloud computing. In AGU Fall Meeting Abstracts (Vol. 2022, pp. S52A-08).
33. Clements, T. (2023). Earthquake detection with tinyml. Seismological Society of America, 94(4), 2030-2039.
34. Mousavi, S. M., & Beroza, G. C. (2023). Machine learning in earthquake seismology. Annual review of earth and planetary sciences, 51, 105-129.
35. Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006, August). Model compression. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 535-541).
36. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
37. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
38. Mousavi, S. M., Sheng, Y., Zhu, W., & Beroza, G. C. (2019). STanford EArthquake Dataset (STEAD): A global data set of seismic signals for AI. IEEE Access, 7, 179464-179476.
39. Swiss Seismological Service (SED) At ETH Zurich. (1983). National seismic networks of Switzerland. ETH Zürich.
40. Xiao, H., Walter, J. I., Ogwari, P., Ho, L. M., Thiel, A. D., Gregg, N., ... & Woelfel, I. (2026). Transfer learning and benchmarking for induced seismic event detection: Insights from Oklahoma. Journal of Geophysical Research: Machine Learning and Computation, 3(4), e2025JH001194.
41. Standard, O. A. S. I. S. (2014). MQTT version 3.1. 1. URL http://docs. oasis-open. org/mqtt/mqtt/v3, 1, 29.
42. Allen, R. V. (1978). Automatic earthquake recognition and timing from single traces. Bulletin of the seismological society of America, 68(5), 1521-1532.
43. Zhao, M., & Chen, S. (2021). The generalization ability research of deep learning algorithm in seismic phase detection of regional seismic network. Earthquake, 41(1), 166-179.
44. Li, J., Zhai, H., Jiang, C., Wang, Z., Wang, P., Chang, X., ... & Si, Z. (2025). Application of artificial intelligence technology in the study of anthropogenic earthquakes: a review. Artificial Intelligence Review, 58(5), 155.

45. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). Pmlr.
46. Zhu, W., Biondi, E., Li, J., Yin, J., Ross, Z. E., & Zhan, Z. (2023). Seismic arrival-time picking on distributed acoustic sensing data using semi-supervised learning. Nature Communications, 14(1), 8192.
47. Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., ... & Soto, H. (2022). SeisBench—A toolbox for machine learning in seismology. Seismological Society of America, 93(3), 1695-1709.
48. Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1314-1324).
49. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.
50. Buciluă, C., Caruana, R., & Niculescu-Mizil, A. (2006, August). Model compression. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 535-541).