SAND2023-04777C

SnT 2023
CTBT: SCIENCE AND TECHNOLOGY CONFERENCE
HOFBURG PALACE - Vienna and Online
19 TO 23 JUNE

Testing a Paired Neural Network to Characterize Aftershock Sequences

Erica Emry, Brendan Donohoe, Andrea Conley, Rigobert Tibi, and Christopher Young
Sandia National Laboratories, Albuquerque, NM

INTRODUCTION

We tested 6 different Paired Neural Network (PNN) models using 2 analyst-validated aftershock sequences.

The PNN models were originally trained on a noise-augmented, constructed, global earthquake dataset.

METHODS/DATA

We tested the PNN models with 2 different 'Match' and 'Non-Match' criteria, in lieu of more extensive, associated information.

START

RESULTS

By most metrics, scores were relatively low – we place most importance on the F1 and AUC scores.

However, our sub-optimal 'Match' and 'Non-Match' criteria likely affects the low-performance.

CONCLUSION

Overall, the existing PNN models struggled to generalize to the available aftershock datasets.

Fine-tuning the PNN models on aftershock data indicates improvement, and provides some ideas for our future directions.

Please do not use this space, a QR code will be automatically overlayed

This Ground-based Nuclear Detonation Detection (GNDD) research was funded by the National Nuclear Security Administration, Defense Nuclear Nonproliferation Research and Development (NNSA DNN R&D).

P3.5-348

SnT 2023
CTBT: SCIENCE AND TECHNOLOGY CONFERENCE
HOFBURG PALACE - Vienna and Online
19 TO 23 JUNE

Introduction: Exploration of Paired Neural Networks to Improve Aftershock Identification

Sandia National Laboratories
U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration

Large magnitude earthquakes & aftershocks sequences unexpectedly occur & greatly increase analyst workload

Although cross-correlation techniques can identify similar earthquakes (like aftershocks):

- **Creation of quality template libraries for in progress sequences can be difficult**
- **Some regions have no historical seismicity to use as templates**
- **Cross-correlation can be affected by spikes and overlapping earthquakes**
- **Cross-correlation tends to be computationally intensive**

Can ML Improve Aftershock Labeling?

Like cross-correlation, if similar events (such as aftershocks) can be rapidly labelled a match or non-match, then it could help alleviate analyst burden and would allow the analyst to maintain attention on other global events.

Any ML model *must* generalize to data from other regions that it wasn't trained on, because:

- The location of the next large magnitude earthquake is unpredictable and could occur in many different regions around the globe.
- We have few recorded earthquakes on some faults that can produce large-magnitude earthquakes (Example: currently locked faults, like the Cascadia subduction zone)

A successful ML model should not mislabel events we care about as an aftershock. Model should have low likelihood of false positives (false classification as aftershock).

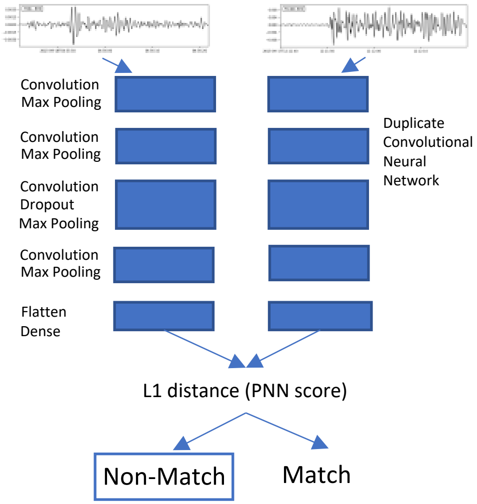
Paired Neural Network (PNN) Models

Conley et al. (2021) trained a PNN model to identify waveform similarity. PNNs (also referred to as Twin or Siamese Neural Networks) are frequently used for image recognition.

Each PNN has 2 branches with the same (duplicated) architecture, and each branch is a convolutional neural network (CNN). Each model has 4 convolutional blocks described below, and a 5th block with a flatten and a dense layer.

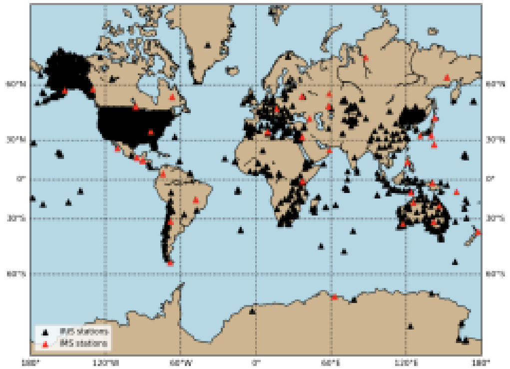
Each convolutional block has 2 or 3 transformations. The convolutional transformation utilizes some number of filters, but the size of the output is not altered. The max pooling transformation takes the top number in each subsample of the waveform, and the size of the output is decreased. The spatial dropout layer removes some number of neurons in the layer to prevent overfitting.

A Monte-Carlo dropout procedure is also used to quantify uncertainty.

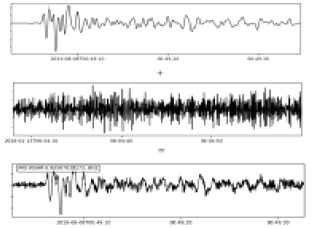


Initial Training Data: Real Event Data, Augmented with Noise

- The training data set was composed of global seismicity, not aftershock earthquakes. This included 15,764 earthquakes and 827 stations.
- Global distribution of available stations was much denser in U.S., Europe, Australia, and East Asia
- Earthquakes from IMS stations spanned 2007-2020, and earthquakes from IRIS stations occurred during 2011
- Noise datasets included the STEAD noise dataset (Mousavi et al., 2019) and the University of Utah noise dataset (Tibi et al., 2021)
- Some waveforms had copies of the event later in the sequence – to simulate overlapping events (common in aftershocks)



Above Figure: Conley et al., (2021) Training Data Station Distribution

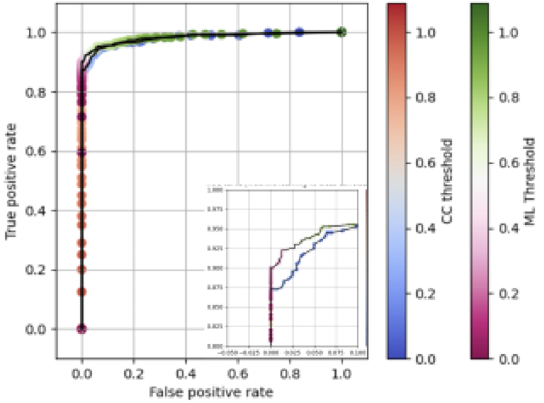


Left Figure: Conley et al., (2021) Constructed waveform (no overlapping events)

Original PNN Model Test Results

- Tested against subset (15%) of constructed data. Training data accounted for 70%, and validation data accounted for 15%.
- The test data was pulled out of full dataset prior to training, and it was selected randomly.
- The PNN models were found to outperform cross-correlation in the top left corner of the ROC curve

Below Figure: Conley et al. (2021) ROC curve comparison w/CC scores



Please do not use this space, a QR code will be automatically overlayed

P3.5-348

Objectives: Explore Generalizability of PNN Models when Tested on Validated Aftershock Sequences

Sandia National Laboratories
U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration

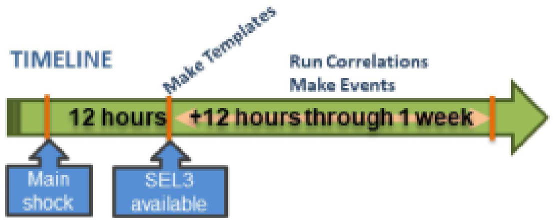
Can ML Improve Aftershock Labeling?

The initial PNN model was trained using a "constructed" dataset. Using global (non-aftershock) earthquakes and various noise libraries to augment the waveforms, pairs of matching and pairs of non-matching time-series were constructed (Conley et al., 2021).

The PNN models were successful when tested against a subset of the constructed database (Conley et al., 2021).

However, the model must generalize to datasets that it has not been trained on – specifically we are interested in applicability to aftershock sequences.

To test this, we utilize arrival datasets of 'true positives' and 'false alarms', corresponding to real aftershock sequences that were previously explored in a cross-correlation study from Sundermier et al. (2019) and validated by an expert analyst. The original cross-correlation detections from Sundermier et al. (2019) were based on templates that were automatically generated from the SEL3 catalog on 12 or 13 different IMS stations, located at regional to teleseismic distances from the sequence.



Sundermier et al. (2019) – Timeline for selecting templates and for running SeisCorr detection software

The two aftershock sequences validated by an expert analyst were the 2015 Illapel, Chile and the 2015 Gorkha, Nepal.

The mainshock magnitude for the 2015 Illapel, Chile event was Mw 8.2. Sundermier et al. (2019) identified 88 template events with 441 distinct template arrivals from the SEL3 catalog. This led to 960 detections on the 12 IMS stations used.

The mainshock magnitude for the 2015 Gorkha, Nepal event was Mw 7.8. Sundermier et al. (2019) identified 91 template events with 353 distinct template arrivals from the SEL3 catalog. This led to 968 detections on the 13 IMS stations used.

These results from Sundermier et al. (2019) were validated by an independent analyst. When comparing the data to the IDC LEB catalog, the analyst marked the arrivals as:

- "True Positive" – In this case, the arrival detected by Sundermier et al. (2019) matched the LEB and was a valid event.
- "Valid Added" – In this case, the arrival detected by Sundermier et al. (2019) was found to be a valid arrival from a valid event that was not in the LEB catalog.
- "False Alarm" – In this case, the arrival detected by Sundermier et al. (2019) was not a valid arrival from a valid event or it was an arrival from non-aftershock earthquake elsewhere in the world.

To Answer the Objective Question: "Can ML Improve Aftershock Labeling?"

We utilize 6 different PNN models trained by constructed (noise-augmented) datasets of global seismicity. They were trained on data that either included or did not include overlapping waveforms, and they were trained at 3 different frequency bands (raw, bandpassed at 1.5-5 Hz, or highpassed at >0.3 Hz).

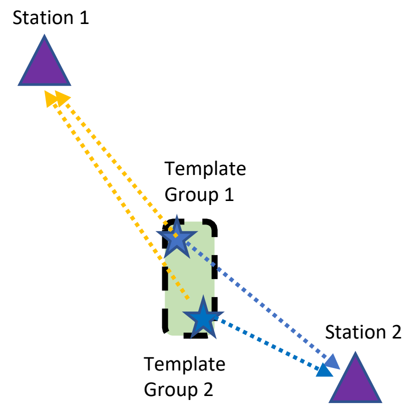
We test the generalizability of these 6 different PNN models on expert-validated datasets for the 2015 Illapel and 2015 Gorkha aftershock sequences.

INTRODUCTION
OBJECTIVES
METHODS/DATA
RESULTS
CONCLUSION

Methods: Determining Ground-Truth and Metrics

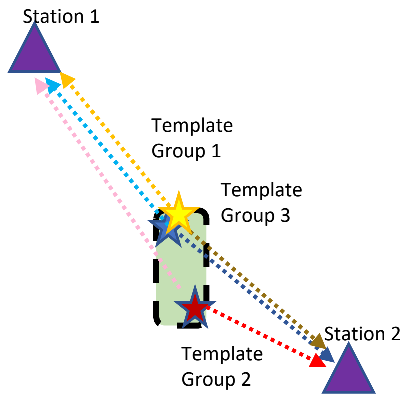
Match/Non-Match Criteria (Ground-Truth):

We leveraged 2 expert-validated aftershock test datasets



Station-Specific, Template-Invariant:

- Benefits:
  - All aftershock arrivals at the same stations are assumed to match
  - Easy to implement, results in more matches
  - Would be better for eventual operational setup
- Drawbacks:
  - Aftershocks from opposite ends of rupture would be called a match, but might not match well (as suggested for template groups 1 and 2, on left)
  - Aftershocks of different source types (normal, thrust, etc.) would be called a match, but might not match well.



Station-Specific, Template-Specific:

- Benefits:
  - All aftershock arrivals detected by the same template assumed to match
  - Aftershocks from opposite ends of rupture are not a match
  - Aftershocks from different source types are not a match
- Drawbacks:
  - Aftershocks detected by nearby (but different) templates might actually be similar and should be a match, but would not be considered as one (as in the case of the gold and blue template groups, on left).

Metrics for Test Results:

Depending on test datasets, different scores can be used to determine model performance

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Requires evenly sized match & nonmatch populations (no "class imbalance")
In our dataset, we have many more non-matches & the PNN models performed best at determining non-matches. So, accuracy would give artificially high values (making it look like our PNN model is better than it is)

$$Precision = \frac{TP}{TP+FP}$$

Model Precision is only sensitive to match predictions
If our 2nd match criteria ("template-specific") leads to many FP in comparison to the 1st match metric ("template-invariant"), we'd expect a decreased precision.

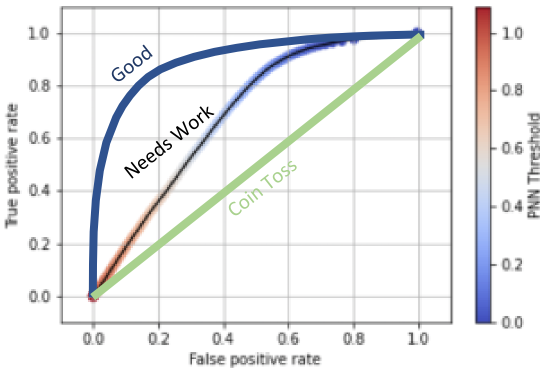
$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{TN+FP}$$

Information for TPR (recall) & FPR with different PNN score threshold assumptions – directly informs the ROC curves

$$F1 = 2 * \frac{Precision * TPR}{Precision + TPR}$$

F1 is the harmonic mean of precision & recall (TPR)
It is specifically designed to balance datasets with class imbalances, like ours.



As PNN score threshold changes (between match & nonmatch), the TPR (Recall) and FPR changes.

A concave downward curve is desired, and results in a high Area-Under-the-Curve (AUC)

Coin Toss: AUC = 0.5
Needs Work: AUC = 0.65
Good: AUC ~ 0.9

Test 5: Bandpass-filtered, Trained without overlapping waveforms, Template-Invariant criteria

INTRODUCTION
OBJECTIVES
METHODS/DATA
RESULTS
CONCLUSION

Please do not use this space, a QR code will be automatically overlayed

Results: Initial Models Struggle to Generalize to Aftershocks, Given Assumed Match/Non-Match Criteria

Sandia National Laboratories
U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration

Models trained with Overlapping Data

Match Criteria #1

Station-Specific & Template-Invariant

(All aftershocks are Matches)

Match Criteria #2

Station-Specific & Template-Specific

(All aftershocks associated to same Templates are Matches)

Test # (Filter)	TPR-to-FPR	AUC	TP	FP	TN	FN	TPR	FPR	Precision	F1
Chile, Station-Specific, Template-Invariant, With Overlapping Signals										
1 (R)	1.20	0.486	606	6738	417899	31247	0.019	0.016	0.083	0.031
2 (BP)	2.05	0.545	2149	13993	410644	29704	0.068	0.033	0.133	0.090
3 (HP)	1.20	0.504	979	10905	413732	30874	0.031	0.026	0.082	0.045
Chile, Station-Specific, Template-Specific, With Overlapping Signals										
7 (R)	1.73	0.515	75	7567	461702	2620	0.028	0.016	0.010	0.015
8 (BP)	3.21	0.587	292	15850	438036	2313	0.112	0.035	0.018	0.031
9 (HP)	1.64	0.528	116	11768	442118	2620	0.042	0.026	0.010	0.016

Bandpass is better (in comparison to raw or highpass)
AUC is higher overall for the template-specific criteria
Clear decrease in Precision in Template-Specific Criteria

Models trained with NO Overlapping Data

Match Criteria #1

Station-Specific & Template-Invariant

(All aftershocks are Matches)

Match Criteria #2

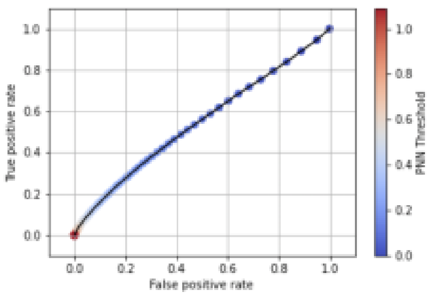
Station-Specific & Template-Specific

(All aftershocks associated to similar Templates are Matches)

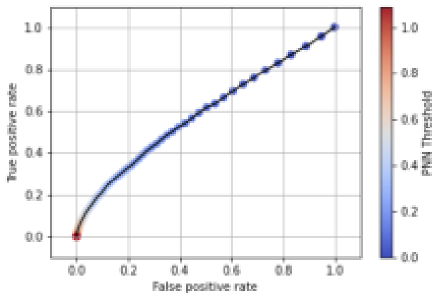
Test # (Filter)	TPR-to-FPR	AUC	TP	FP	TN	FN	TPR	FPR	Precision	F1
Chile, Station-Specific, Template-Invariant, Without Overlapping Signals										
4 (R)	2.98	0.677	4192	18745	405892	27661	0.132	0.044	0.183	0.153
5 (BP)	1.75	0.688	16869	128390	296247	14984	0.530	0.302	0.116	0.190
6 (HP)	2.22	0.692	4867	29215	395422	26986	0.153	0.069	0.143	0.148
Chile, Station-Specific, Template-Specific, Without Overlapping Signals										
10 (R)	2.69	0.654	349	22588	431297	2256	0.134	0.050	0.015	0.027
11 (BP)	1.66	0.687	1357	143304	310598	1230	0.525	0.316	0.009	0.017
12 (HP)	2.32	0.689	447	33635	420250	2158	0.172	0.074	0.013	0.024

AUC is highest overall for these tests
Precision and F1 are highest, assuming Template-Invariant Criteria
Clear decrease in Precision and F1 in Template-Specific Criteria

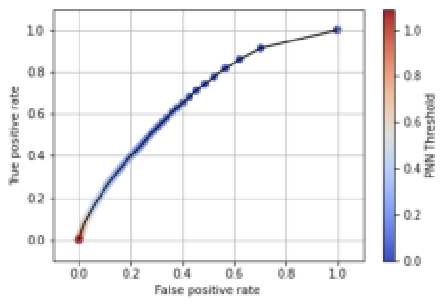
****All numbers (except AUC) are for PNN threshold score = 0.5**



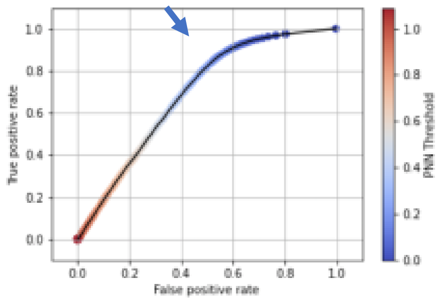
Test 2: Template-Invariant Bandpass, Trained with Overlapping Data



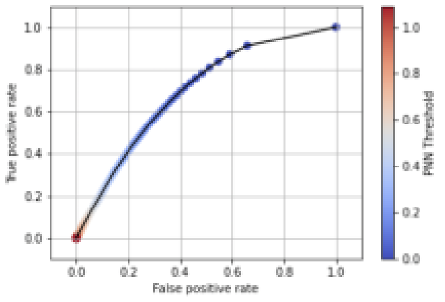
Test 8: Template-Specific Bandpass, Trained with Overlapping Data



Test 4: Template-Invariant Raw, Trained with NO Overlapping Data



Test 5: Template-Invariant Bandpass-Filtered, Trained with NO Overlapping Data



Test 6: Template-Invariant Bandpass-Filtered, Trained with NO Overlapping Data

INTRODUCTION
OBJECTIVES
METHODS/DATA
RESULTS
CONCLUSION

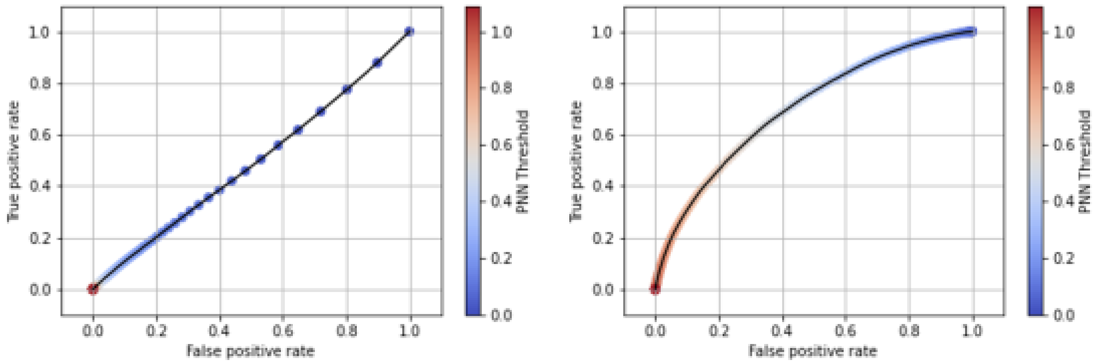
Please do not use this space, a QR code will be automatically overlayed

P3.5-348

Conclusion: Trained PNN Models Struggle with Real Aftershock Sequences, but may Improve with Fine-Tuning

Conclusions and Initial Results from Fine-Tuning Studies

- PNN models trained with constructed data struggled to generalize to the aftershock datasets
- However, the available datasets need more associated information to better quantify the “Match” vs. “Non-Match” criteria
- Initial Results from Fine-Tuning of existing PNN Models with some aftershocks shows improvements
  - However, fine-tuning is not the best answer for rapid response to an in-progress aftershock sequence.



Initial results from fine-tuning of PNN models

Above shows improvement (right) after fine-tuning the original PNN model trained with overlapping, raw waveforms (left)

Future Directions

- We need better criteria for defining a Match vs. a Non-Match
  - Explore template similarity & cross-correlation scores (for all waveform combinations)
- We need to better explore how PNN models are trained
  - What’s different between how original models were trained, and why did tuning with aftershock data improve it?
  - Geographic Distribution of data?
    - New validated 2011 Tohoku aftershock sequence data!
    - Data from Japan was better represented in the training dataset – so maybe the original PNN models will work better here…?
  - Original model used training data that were the same base waveform, but with different amounts of noise added.
    - Are those waveforms not different enough?
- We used a contrastive loss function, but would other loss functions be more appropriate? (e.g. Triplet Loss, as in Dickey et el., 2019)

INTRODUCTION
OBJECTIVES
METHODS/DATA
RESULTS
CONCLUSION

References

References

- Conley, A., B. Donohoe, and B. Greene (2021), Aftershock Identification Using a Paired Neural Network Applied to Constructed Data, *OSTI Technical Report,* doi:10.2172/1821802
- Dickey, J., B. Borghetti, W. Junek, and R. Martin (2019), Beyond Correlation: A Path-Invariant Measure for Seismogram Similarity, *Seismological Research Letters, 91*(1), 356-369, doi:10.1785/0220190090
- Mousavi, S.M., Y. Sheng, W. Zhu, and G.C. Beroza (2019), Stanford Earthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI, *IEEE Access, 7,* 179464-179476, doi:10.1109/ACCESS.2019.2947848
- Sundermier, A., R. Tibi, and C.J. Young (2019), Applying Waveform Correlation to Aftershock Sequences Using a Global Sparse Network, *OSTI Technical Report,* doi:10.2172/1763210
- Tibi, R., P. Hammond, R. Brogan, C.J. Young, and K. Koper (2021), Deep Learning Denoising Applied to Regional Distance Seismic Data in Utah, *Bulleting of the Seismological Society of America, 111*(2), 775-790, doi:10.1785/0120200292

INTRODUCTION
OBJECTIVES
METHODS/DATA
RESULTS
CONCLUSION