Dear Editor,

We sincerely thank you and the reviewers for the careful evaluation of our manuscript. The reviewers' thoughtful and constructive comments have been extremely helpful in improving both the clarity and the scientific rigor of the study.

In the revised manuscript, we have made several substantial updates in response to the reviewers' suggestions. First, we expanded the model comparison by evaluating CNN performance both with and without origin-hour information, and by including three additional seismology-oriented CNN architectures for comparison. We also revised the representation of event origin hour by replacing the original 0–23 encoding with a cyclic sine–cosine formulation, which more appropriately reflects the periodic nature of time-of-day information. These results are now presented in the main text and supplementary material.

We also revised the performance evaluation framework to provide a more comprehensive comparison beyond classification accuracy alone. In particular, we placed greater emphasis on a cost-based evaluation that explicitly accounts for false positives and false negatives, making the model assessment more consistent with standard machine-learning practice and with the practical goals of this study. We further incorporated repeated training and testing with different random seeds during optimization in order to better assess model robustness. In addition, we conducted new experiments to evaluate the influence of different blast-event fractions in the training data.

Most importantly, we developed and applied a new event-level classification framework that converts station-level predictions into event-level probabilities. In this framework, station-specific weights are estimated using logistic regression as a function of both source–station distance and signal-to-noise ratio (SNR). Rather than assigning weights using an arbitrary Gaussian decay function in the original manuscript, the new weighting scheme is learned directly from the training data, enabling it to reflect the region-specific relationship between waveform quality, propagation effects, and classification reliability. We also introduced SNR-based quality-control thresholds and minimum-station requirements to reduce the influence of noisy or weakly constrained observations.

In addition, we revised the construction of the training, validation, and test datasets by adopting a fixed SNR threshold of 1.5 (approximately 1.8 dB), which increased the total number of usable waveforms by about 20,000. Despite this expansion of the dataset, the overall classification performance remained strong and did not decrease substantially.

Please refer in particular to the revised Results section, especially the model-comparison and classification-framework subsections, as well as the Application section, for full details of these improvements.

In addition to these major revisions, we also addressed the reviewers' minor comments and made further improvements to the manuscript's organization, wording, and grammar throughout.

We believe that these revisions have significantly strengthened the manuscript, and we appreciate the opportunity to resubmit it for consideration.

In the following pages, we provide a detailed, point-by-point response to all reviewers' major comments.

Thank you for your consideration.

Sincerely,

Justin Chien and Yajing Liu

Reviewer 2

Major Issues:

1. Line 138 - 149 - What about a standard bandpass filter? Now each signal is treated differently by the DeepDenoiser. Hence, the data becomes more heterogeneous. In the given example in Figure 4, there is a sharp decrease in the coda part of the subplot f, which is, most probably, the result of the DeepDenoiser. Do you have such extreme filtering results in other quarry-blast waveforms?

   Our primary objective is to enhance the time–frequency representation of seismic signals for spectrogram-based classification, rather than to preserve the full waveform characteristics for detailed physical interpretation. A 1 Hz high-pass filter is applied to remove long-period trends and low-frequency noise, which stabilizes the spectrograms and enhances the visibility of the high-frequency features most relevant for discriminating earthquakes from blasts.

   We also tested an additional 1–40 Hz bandpass filter and found that it did not significantly change the resulting spectrogram patterns or the distinction between

earthquake and blast signals. We therefore retained only the high-pass filter to avoid imposing additional assumptions on signal bandwidth and to preserve as much of the original frequency content as possible.

Regarding DeepDenoiser, we agree that denoising can introduce waveform-dependent modifications. In particular, the sharp reduction in the coda observed in Fig. 4f is likely related to the denoising process. Similar behavior is also shown in the official SeisBench DeepDenoiser example, in which coda energy is suppressed to reveal an embedded microearthquake signal. After examining additional quarry-blast waveforms, we found that such strong coda attenuation occurs mainly in low-SNR cases, such as smaller-magnitude events (<2) or events recorded at source-to-station distances greater than 100 km, and is not representative of most of the dataset.

Importantly, the purpose of denoising in this study is to improve the clarity and discriminability of time–frequency features in the spectrograms, rather than to preserve absolute waveform amplitudes or coda characteristics. The classification is based on the overall time–frequency patterns, such as the more vertically concentrated energy commonly observed for earthquakes versus the more horizontally distributed energy often seen for blasts, rather than on their detailed amplitude information. We have revised the manuscript to clarify that the preprocessing steps were designed specifically to improve the distinction between these source types in the time--frequency domain for classification purposes. (Lines 178–180 in revised manuscript)

2. Line 62 - The ML detection tool can detect quarry blast peaks because P-waves exhibit similar features. Near-blast stations may lack S-wave and monochromatic wavefield, while far-blast stations may record S-wave-like signals (as explained in Line 77-84). The accuracy of ML pickers trained on tectonic events in quarry blasts depends heavily on the source-station distance. Please clearly state it in the Introduction.

   We have revised the Introduction to state earlier and more explicitly that ML phase pickers trained on tectonic earthquakes may also detect quarry blasts, and that this behavior can depend strongly on source–station distance and propagation effects. This point is now introduced in the motivation section and followed by a more detailed discussion later in the text.

3. Lines 90-92 - Quantifying the unsuccessful attempt can underscore the importance of the ML-based algorithm developed in this study.

In the revised manuscript, we now explicitly report the performance of directly applying the previous CNN model to Eastern Canada, including that the classification accuracy for pre-labeled blast events is below 90% . This addition better illustrates the limitations of direct model transfer and more clearly motivates the need for the region-specific workflow developed in this study. (Lines 107–199 in revised manuscript)

4. Line 178-179 - This may not be the most appropriate form of data augmentation for seismic signals, since the recorded waveforms cannot be physically rotated in a way that is rotated in the paper. An alternative could be to move the signals back and forth within the 120s-long input to ensure that the P and S waves are not located at the same parts of the spectrum.

   In this study, data augmentation was applied to spectrogram patterns rather than to raw seismic waveforms. The CNN operates on time–frequency representations, and small rotations (1–10°) and horizontal flips were used as standard image-based regularization techniques to improve generalization and reduce overfitting. These operations do not represent physical rotation of seismic signals in the time domain; instead, they introduce minor geometric variability in the image domain, as is commonly done in image-classification tasks.

   The applied rotations were intentionally limited to small angles in order to avoid distorting the fundamental time–frequency structure, including the relative positions of P- and S-wave energy bands. The purpose is not to alter the physical characteristics of the signal, but to improve the model's tolerance to small variations in spectrogram scaling and localization.

   We agree that temporal shifting within the 120 s window is also a reasonable augmentation strategy. In practice, some variability in phase-arrival position is already present because of origin-time uncertainty and differences in source–station distance. We have clarified this rationale in the revised manuscript (Lines 238–242 in revised manuscript).

5. Lines 182-188 - An earlystop parameter can be introduced to avoid overfitting.

   We have now incorporated early stopping throughout the model testing and training workflow to reduce overfitting and improve consistency in model selection.

6. Line 237 - In Lines 127-129, the authors estimated the theoretical P- and S-wave arrivals. However, in Section 4, the authors use actual P- and S-phase picks, which can be tricky because near-fault S-waves are very hard to detect, as the

authors also note in Lines 78-79. Can authors explain the data processing step for Section 4 in more detail?

The theoretical P- and S-wave arrival times described in first submission of Lines 127–129 were estimated for NEDB catalog events, including both blasts and earthquakes, solely for defining signal and noise windows used in SNR calculation. These theoretical arrivals were not used for event detection or phase association. Based on these SNR measurements, waveform denoising was then applied, and only waveforms with SNR > 1.5 within a 1.5° source–station radius were retained for subsequent analysis.

By contrast, Section 4 describes a separate catalog-enhancement workflow based on ML phase picking and phase association. At that stage, event candidates are identified from automatically detected P- and S-phase picks rather than from theoretical arrivals. We agree that S-wave identification can be difficult at near-source stations, as noted in the manuscript. However, the phase-association procedure uses observations from multiple stations, so even if S-wave picks are unclear or absent at near stations, they may still be detected at more distant stations where the signal is more distinct. Candidate events are retained only if they satisfy the association criteria, including a minimum of eight total picks with at least three P and three S phases.

We have revised the text to make this distinction clearer and to explain in greater detail that theoretical arrivals are used only for SNR estimation in the training-data construction (Lines 171–177 in revised manuscript), whereas actual ML-derived picks are used in the catalog-enhancement workflow (Text S1).

7. Line 300 - Authors have to deal with the heterogeneity of the sampling rates in any situation. Moreover, the USArray was a temporary network, while the Canadian Networks are permanent. In any case, I believe, the system has to be optimized for the permanent ones.

   We agree that heterogeneity in sampling rate and network configuration must be addressed carefully in any large-scale catalog-enhancement framework.

   Although USArray was a temporary deployment, it played a major role during its period of operation in Eastern Canada. While USArray stations were active in the study region, station density was substantially higher than during periods covered only by the permanent Canadian networks. This denser coverage improved detection capability and waveform sampling, resulting in a much larger set of detected events in the enhanced catalog that required source classification.

For this reason, including waveform data from the USArray period in the training dataset is important for building a robust classifier. It allows the model to learn from a denser and more informative observational setting and improves generalization across a wider range of recording conditions. Moreover, although USArray was temporary, its coverage in Eastern Canada lasted for approximately two years, representing a scientifically valuable interval rather than a short-lived experiment.

At the same time, we agree that long-term monitoring ultimately depends on the permanent Canadian networks. Our framework is applicable to the current CNSN permanent stations, and further optimization specifically tailored to regional/variable network configurations is an important direction for future work. We have revised the manuscript to clarify this balance between the value of incorporating USArray data during training and the long-term applicability of the framework to permanent monitoring networks. (Lines 211–220 in revised manuscript)

8. Figure 2 - ML Enhanced picks are quite larger with respect to the NEDB catalog. Is there any confidence level information about the ML-picked phases? Mispicking can be a problem, or some other anthropogenic sources may be mispicked as event signals. Can the author explain if there was any extra step to check the data quality of the newly detected picks?

   We agree that the initial ML-enhanced catalog may contain false detections arising from mispicked phases, noise transients, or other anthropogenic sources. In the revised manuscript, we now clarify that an additional quality-control step was applied after phase association: signal-to-noise ratio was used as a first-order screening criterion to remove likely noise events before source classification. This procedure is now described in the main text in Section 4.1 for the 2020–2022 WQSZ application.

   Even with this screening, we acknowledge that the initial ML-enhanced catalog still contains a mixture of true events and false detections. A more exhaustive validation of all newly detected events would be valuable, but such a detailed catalog-verification effort is beyond the scope of the present study, which focuses primarily on source discrimination between earthquakes and blasts after the initial catalog-enhancement stage.

   In a separate study focused specifically on WQSZ catalog enhancement, using the original initial enhanced catalog generated with the default parameters of EQTransformer and PyOcto phase association, a random check of 100 events

showed that approximately 20% were false detections. In the present study, we did not perform visual inspection for every detection.

9. Figure 5 - Also in the text, they are not really RGB images, I assume, but the EW, NS, and Ver. components. Use the analogy (RGB) to explain the data to the reader once, then continue with the term component.

   We have revised both the figure caption and the main text (Lines 209–210 in revised manuscript) to clarify that the input consists of three-component waveform spectrograms (EW, NS, and vertical), arranged analogously to the three channels of a standard RGB image.

10. Figure 6 - Learning rate affects ConvNext and VGG-16, while others are not susceptible to the parameter. Apart from that, all models perform roughly the same in terms of accuracy. Those models are probably quite complex, and hence they achieve very good results. Since the authors mentioned training a model without powerful hardware, would it be better to develop simpler AI architectures to understand how much simplicity is not enough to achieve similar accuracy rates (eg. https://doi.org/10.1007/s00024-024-03440-0)? Moreover, previously developed models can also be tested (eg. https://doi.org/10.1785/0120240244).

    To address this point, we have expanded the model comparison by testing three additional simpler CNN architectures designed specifically for classification. This broader comparison allows us to better assess the trade-off between model complexity and performance, and to evaluate how much simplification is possible while still maintaining strong classification capability. The revised manuscript now includes these additional results and corresponding discussion. Overall, the simpler CNN architectures consistently performed worse than the more complex image-classification models, regardless of whether the event origin-time feature was included. These results indicate that, while simpler models may be more attractive from a computational perspective, the more complex architectures remain better suited for achieving the highest classification performance in this application.

11. Figure 7 - In the Figure, ML-based Blasts cover a large area, and there are clusters of quarry blasts where no manually labelled blasts are detected. Did the authors check via satellite images to see if there are really open quarries or any signs of underground mining operations (if it is still a valid form of mining in Canada)? Figure 9 - There are some quarry blasts detected on the St. Lawrence River where no manually labeled blasts are available. This raises questions about the reliability of the quarry blast classifications, which would benefit from further validation and a more explicit discussion in the Discussion section.

We agree that these spatially isolated or weakly validated blast clusters require careful interpretation. We did not carry out a systematic satellite-image survey of all predicted blast locations to verify the presence of open quarries or underground mining activity. Such validation would be valuable, especially for areas where no manually labeled blasts are available nearby.

In the Montreal and Ottawa regions, however, many of the ML-classified blasts are close to blast events identified in the NEDB public catalog, although some spatial scatter remains. In regions such as Charlevoix and along parts of the St. Lawrence River, the interpretation is less certain, and these cases may reflect location uncertainty, incomplete blast labeling in the NEDB reference catalog, or misclassification. We also note from personal communication with NRCan seismologists that the NEDB public blast event catalog is often incomplete; NRCan keeps an internal catalog of suspect blasts that require further verification. For example, in another study [Liu et al., revised, Seismica] our group used combined on- and offshore data to detect seismic sources in the Lower St. Lawren Seismic Zone (further downstream from Charlevoix). Of the total 31 detections in October 2023, 8 are labeled as suspect blasts in the NEDB internal catalog but none in the public catalog. We have revised the Discussion to address these uncertainties more explicitly and to note that further external validation, including satellite or industrial land-use information, would be useful in future work.

Reviewer 1

Major comments:

1. The learning database is composed of signals recorded with a sampling rate of 40 Hz and after 2016 (end of the US Transportable array) the seismograms SR is at 100 Hz. In order to use the same frequency range the 40 Hz signals are resampled. Such an operation gives dark blue images between 20 and 50 Hz because no HF energy has been recorded, and we can imagine that a training with such ‘blind’ areas will not be relevant for small magnitude EQ or blast signals that have both high frequency contents. The main problem of Linville et al. (2019) study is also a question of frequency range. Low frequencies are no relevant for a discrimination of high frequency events (roughly between 2 and 100 Hz).

   We agree that combining waveform data recorded at different sampling rates introduces an important limitation for spectrogram-based classification. In particular, when 40 Hz records are resampled to 100 Hz, no additional physical information is created above the original Nyquist frequency of 20 Hz. As a result,

the upper-frequency portion of those spectrograms does not contain true high-frequency signal content.

Our intent, however, is not to rely on artificially created high-frequency information, but rather on the overall morphology of the spectrogram. For resampled earthquake records, the characteristic pattern still appears as a relatively vertically concentrated band, although its frequency extent may be shorter and lower because energy above 20 Hz is not physically resolved. Similarly, blast spectrograms generally retain their more horizontally (time-axis) extended pattern. In this sense, the classifier is designed to distinguish source-dependent spectrogram image patterns rather than depending directly on signal energy in the extrapolated frequency range.

At the same time, we acknowledge that this sampling-rate heterogeneity may reduce the representation of very high-frequency features, especially for small-magnitude earthquakes and blasts. We have revised the manuscript to clarify this limitation and to explain that the use of a common spectrogram size was primarily intended to ensure consistent CNN input dimensions across the full dataset. (Lines 211–220 in revised manuscript)

2. The fact that the explosion time is necessary to improve the CNN is somewhat surprising. I would have suggested exactly the opposite: it is a good indicator for checking that the CNN is working correctly. If all the discriminated explosions occur during opening hours, that seems logical.

   We agree that event origin time should not be treated as a standalone discriminator, nor should it be used as the sole basis for validating whether the CNN is performing correctly. In this study, origin time is used only as an auxiliary feature that complements the waveform-based information extracted from the spectrograms.

   Across all CNN models tested, classification performance improved when origin time was included as an additional feature. This likely reflects the fact that quarry blasts in Eastern Canada occur more frequently during daytime working hours. However, a non-negligible number of blast events also occur during evening, nighttime, or near-midnight hours. For this reason, origin time alone is neither sufficient for robust discrimination nor appropriate as a standalone validation metric.

   New figure 6 shows that incorporating event origin time consistently improves overall classification performance across all tested architectures. This pattern is observed not only for the more complex CNN image-classification models, but

also for the simpler CNN structures developed specifically for seismic signal discrimination. Although the magnitude of improvement varies among models, the inclusion of origin-time information leads to better overall results in every case, indicating that it provides useful supplementary information when combined with waveform-based spectrogram features.

3. The idea of a weight which depends on epicentral distance is risky because it appears that it is sometimes easier to discriminate farer events since wave packets are more distinguishable due to propagation effect and the velocity ratio between P and S waves. Finally the choice of a sum of each station individual score to produce a final score at the network level can be discussed because, at the end, it is good to have results close to 1 or 0. Figures 8 and 9 show that many event are not easy to be labeled which is a sign that the CNN does not work properly. Confusion matrices do not take this subtlety into account. They report classification using a binary view, whereas this is not always the case.

   We agree that the relationship between classification reliability and epicentral distance is not necessarily monotonic. In some cases, as the reviewer pointed out, more distant records may be easier to interpret because propagation separates the wave packets more clearly and increases the apparent P--S time separation. For this reason, in the revised manuscript we no longer use the arbitrary Gaussian distance-decay weighting adopted in the original version.

   Instead, we developed a new event-level classification framework (Section 3.3 in revised manuscript) in which station-specific weights are estimated empirically from the training data using logistic regression as a joint function of source--station distance and denoised SNR. This data-driven approach does not assume that near-source stations are always more reliable. Rather, it allows station reliability to vary according to the combined effects of distance and signal quality, such that even more distant stations may receive relatively high weights when their denoised SNR is high and the training data indicate strong classification reliability under those conditions. We also introduced SNR-based quality-control thresholds and minimum-station requirements to reduce the influence of noisy or weakly constrained observations before event-level classification.

   Regarding the combination of station-level predictions, our objective was not to force the final event-level probability toward 0 or 1, but to preserve the probabilistic information from multiple stations while giving greater influence to records expected to be more reliable. In this context, intermediate event-level values are meaningful and should not be interpreted simply as evidence that the

CNN is failing. Rather, they reflect cases in which the available station evidence is mixed or ambiguous, even after low-quality observations have been excluded through SNR screening.

We therefore agree that confusion matrices alone do not fully capture this probabilistic subtlety. In the revised manuscript, we clarify that event-level outputs should be interpreted probabilistically rather than purely in a binary sense. In addition, the complete event-level classification workflow, including threshold selection, was evaluated across 10 independent random seeds. Intermediate prediction values therefore remain informative, as they represent uncertainty in the available station evidence rather than necessarily poor model performance.

4. Color scale for spectrograms: if spectrograms are not in black-and-white or monochrome then it can be a problem for an efficient learning stage due to energy report from one component to another one. Such features are less visible in gray scale spectrograms compared to the navy-to-yellow color scale images.

   In our workflow, the model input does not consist of RGB images generated from a plotting colormap. Instead, each sample is a three-dimensional spectrogram array in which the three channels correspond directly to the three seismic components (north–south, east–west, and vertical). Each component is represented by its own normalized spectrogram matrix, and the resulting 129×92×3 array is then resized to 224×224×3 before being passed to the CNN.

   The model therefore learns from standardized component-specific spectrogram values rather than from artificial color contrasts introduced by a display colormap. The navy-to-yellow color scheme appears only in the manuscript figures for visualization and conceptual illustration, and is not used as part of the model input. We have revised the text to make this distinction clearer (Lines 209–210 in revised manuscript).

5. What is the law concerning authorized times for industrial blasts? When looking at figure S6, it seems that few hundred of blasts (in the period 2000-2024) are occuring during the night is it possible? How much confidence can we have in the blast database?

   The figure has now moved to figure S3.

The blast labels in the public NEDB catalog are based on manual review by analyst(s), and we therefore treat them as the best available reference labels for this study. Most labeled blasts occur during expected daytime working hours, but a smaller number are listed during evening or nighttime periods. These cases may reflect blasting activity outside typical hours, although we cannot independently verify the operational circumstances of each individual event.

At the same time, the public NEDB blast catalog is not intended to be a complete record of all industrial blasts. Based on communication with NRCan, earthquakes are the primary cataloging priority, with non-blast anthropogenic events representing a secondary priority, and blasts are therefore not comprehensively included in the public database. As a result, the public catalog should be regarded as a useful but incomplete reference for blast events. We have revised the manuscript to clarify that the blast labels used here represent the best available analyst-reviewed reference data, but that the public blast database is likely incomplete.

6. Is it possible to describe how the blast database is made? Based on what? Visual ispection of seismograms? Is there any feedback possible in the terms of wrong labeled events?

   The NEDB blast catalog used in this study is based on manually reviewed events identified by NRCan analysts. In addition to these labeled blast events, the NEDB internal catalog shared from their analyst also contains many unlabeled events that may potentially include additional blasts but were not assigned an explicit blast label.

   We agree that some labeled events may still be uncertain or potentially misclassified, and that feedback on incorrect labels is valuable. One potential contribution of our classification model is to help identify events whose waveform characteristics are strongly consistent with blasts but are not labeled as such in the existing catalog. Such cases could provide useful candidates for further analyst review and future catalog refinement.

7. The same amounts of quakes and blasts are used for training and validation stages which is a good thing, but do authors tested what's happen if there is a large difference in the two datasets?

Yes. In the revised study, we added an additional set of experiments to evaluate the effect of class imbalance between earthquake and blast samples, with blast fractions ranging from 0.05 to 0.5 relative to the number of earthquake samples. These tests were designed to assess how sensitive the model performance is to differences in class proportions and to verify that the main conclusions are not dependent on using a strictly balanced dataset. We now report these results in the revised manuscript Discussion section Line 450-458 and Fig.S18.