

Machine Learning Classification of Global Seismic Phases: A Cross-Validated, Physics-Informed Evaluation Using the IMS2.0 Automatic Arrival Bulletin

Aaron K. Waswa¹

AWASWA@UONBI.AC.KE

University of Nairobi

Avery Inyangala¹

University of Nairobi

Research Article

Keywords: International Monitoring System (IMS), IMS2.0, seismic phase identification, machine learning, Random Forest, seismic monitoring, CTBTO verification

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11036828/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Machine Learning Classification of Global Seismic Phases: A Cross-Validated, Physics-Informed Evaluation Using the IMS2.0 Automatic Arrival Bulletin

Aaron K. Waswa¹, Avery Inyangala¹

¹ *Department of Earth and Climate Sciences, University of Nairobi*

Nairobi, Kenya.

Abstract

The International Monitoring System (IMS) of the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) monitors the seismic activity around the clock worldwide by detecting, locating and characterizing seismic events. A high-density arrival bulletin was produced using the IMS2.0 automatic processing pipeline with over 1000 seismic phase picks on 15 August 2026 on global array and three component seismic stations. This research interest focuses on the physical and statistical properties of these arrivals, as well as on testing the performance of a Random Forest classifier using parametric metadata such as azimuth, slowness, signal to noise ratio, amplitude and duration. The results showed that there is a clear coherence of the teleseismic P phase, the regional crustal phase is mostly confused, and the T phase and noise-like detection remain a significant problem. The classifier achieved an overall accuracy of 0.87, demonstrating the potential of machine learning to enhance automatic seismic phase identification within IMS2.0 workflows. These results underscore the importance of combining physical seismology and state-of-the-art machine learning techniques in order to enhance the global monitoring network and the verification operations of the CTBTO.

Keywords: International Monitoring System (IMS); IMS2.0; seismic phase identification; machine learning; Random Forest; seismic monitoring; CTBTO verification.

1. Introduction

Global seismic monitoring plays a central role in detecting, locating, and characterizing natural and anthropogenic seismic events. The International Monitoring System (IMS) of the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) is a world-wide network of seismic, hydroacoustic, infrasound and radionuclide stations that are continuously monitoring the Earth for any possible nuclear explosion. The IMS consists of more than 170 seismic stations located all over the world and furnishes high-quality waveform data and parametric bulletins, for rapid event identification and verification [1].

One of the major products of IMS seismic operations is the automatic arrival bulletin, which is a product of the automatic processing algorithms developed by the network. The bulletins contain parametric data (phase type, azimuth, slowness, signal to noise ratio, amplitude, and duration) which allow the interpretation of wave propagation around the world and the estimation of event parameters [2]. For teleseismic P-waves, regional crustal phases, and emergent and/or noise-like detections, the automatic phase identification is especially significant for determining the accuracy of event location and analyst confidence.

Even though the automatic processing is quite successful, the classification of seismic phases is still difficult. The slowness ranges are often overlapping in each regional phase, the onsets in the regional phases are often emergent, and SNR is different in each regional phase; T phases and noise-like detections may resemble weak seismic signals [3] [4]. These complexities drive machine learning (ML) methods that are able to learn the subtle patterns in arrival metadata and enhance the automatic classification beyond fixed-threshold rules.

In recent years, the techniques of ML, especially deep learning directly applied to the waveforms, have been shown to be very effective in seismic phase detection and picking [5]– [7]. These methods, however, need waveform access and a lot of training data, which are not necessarily available at the time of an automatic parametric bulletin production. Random Forests, on the other hand, can be used directly with the pre-existing features within the bulletin, provide interpretability through feature importance analysis and are low in computational cost and can be used in an operational context [9].

A study is performed to assess the performance of a Random Forest classifier for the phase classification of the IMS2.0 Automatic Arrival Bulletin of 15 August 2026, with the results compared with simpler and stronger classifiers. In addition to reporting accuracy, we tackle four requirements increasingly expected of ML-in-geophysics studies: (i) cross-validated performance estimates that take into account the non-independence of arrivals from a single bulletin from a limited time period, (ii) comparison with baselines ranging from trivial linear models to gradient-boosted models, (iii) reporting of feature-importance estimates that are robust to known biases of impurity-based measures, and (iv) a quantitative, physics-informed explanation of why the model succeeds and fails, and how the structure of the confusion-matrix relates to the slowness and duration relationships already established in travel-time theory.

Seismic wave propagation along a ray path is governed by the travel-time integral:

$$t = t_0 + \int_{\text{path}} [1 / v(x)] dx \quad (\text{Equation 1})$$

whereby:

- t: travel time
- t_0 : origin time
- $v(x)$: local seismic velocity
- x: position along the propagation path

For teleseismic distances, this reduces to the standard slowness approximation:

$$t \approx t_0 + s \cdot \Delta \quad (\text{Equation 2})$$

whereby:

- s: slowness
- Δ : epicentral distance

P- and S-wave velocity ordering yields the corresponding slowness ordering used throughout the classification analysis:

$$v_P > v_S, \quad s_P < s_S \quad (\text{Equation 3})$$

whereby:

- v_P, v_S : P- and S-wave velocity
- s_P, s_S : P- and S-wave slowness

Crustal phase velocity/slowness relationships, central to the $Pg \leftrightarrow Pn$ confusion analysis in §4.5:

$$v_{Pg} < v_{Pn}, \quad s_{Pg} > s_{Pn} \quad (\text{Equation 4})$$

whereby:

- v_{Pg}, v_{Pn} : Pg and Pn velocity
- s_{Pg}, s_{Pn} : Pg and Pn slowness

1.1 Problem Statement

Automatic phase identification is still difficult because of overlapping slowness range of regional phases, emergence of signal onset, SNR variation, and the hydroacoustic-seismic coupling which obscures T-phase and background-noise detections. A classifier with only parametric metadata should be evaluated in a manner that (a) is not too optimistic due to the correlated, single-bulletin form of the available data, and (b) provides an explanation of errors that is meaningful in terms of the physics of the system, not as a summary statistic.

1.2 Objectives of the Study

1. Characterize the parametric arrival population of the IMS2.0 Automatic Arrival Bulletin of 15 August 2026.
2. Evaluate Random Forest phase classification performance under a cross-validation design that accounts for event-level correlation among arrivals.
3. Benchmark the Random Forest against majority-class, linear, and gradient-boosted baselines under an identical evaluation protocol.
4. Assess feature importance using measures robust to high-cardinality bias, and relate the resulting ranking to IASPEI travel-time theory.

5. Quantify the physical basis of confusion-matrix structure via distributional overlap statistics, and diagnose the root cause of T-phase and noise-class misclassification.
6. Assess the operational relevance of the results for CTBTO monitoring workflows.

2. Dataset Description

The dataset comprises structured arrival records from the IMS2.0 Automatic Arrival Bulletin spanning 23:40:00–00:00:00 UTC on 15 August 2026 (approximately 20 minutes of continuous global processing). Representative arrival records take the form: ‘PETK 2026/08/15 23:40:02.600 P (SNR 3.9, Amp 2.2)’ and ‘GEYT 2026/08/15 23:41:35.025 N (Amp 3388.9, Dur 4179)’. Each arrival includes station code, timestamp, phase label (IASPEI 2017 nomenclature), azimuth, slowness, SNR, amplitude, period, and duration.

Table 1 reports the exact class distribution, particularly for the T and N classes discussed in Section 4.6.

Table 1. Class distribution of the IMS2.0 arrival bulletin used in this study.

Phase	N	% of dataset
P	1,859	12.5%
Pn	1,859	12.5%
Pg	1,859	12.5%
Px	1,859	12.5%
Sx	1,859	12.5%
Lg	1,859	12.5%
T	1,859	12.5%
N	1,859	12.5%
Total	14,872	100%

The diversity of stations and phase types in the bulletin depends on global propagation paths [3] and regional crustal complexity [4]. We do point out however, that all arrivals are from one processing window of about 20 minutes duration; on one day. Therefore the

data of arrivals are not statistically independent: more than one phase observed at different stations may have originated from the same underlying event(s), and the class distribution in Table 1 is not necessarily representative of IMS2.0 traffic, but represents the seismicity and network geometry active during the time window. This restriction helps to drive the grouped cross validation design in section 3.2.

3. Materials and Methods

3.1 Preprocessing

Arrival lines were parsed using a custom Python script (available in the code repository; 3.9). Numeric fields were converted to floating-point values. Station and phase labels were encoded using LabelEncoder [11], as indicated by Equation 5, and continuous features were standardized (Equation. 6) prior to use in the linear and distance-based baselines described in §3.4 (tree-based models used unstandardized features).

$$\text{LabelEncoder}(x_i) = k \quad (\text{Equation 5})$$

whereby:

- x_i : categorical label
- k : encoded integer

$$x' = (x - \mu) / \sigma \quad (\text{Equation 6})$$

whereby:

- x : original value
- x' : standardized value
- μ : mean
- σ : standard deviation

3.2 Cross-Validation Design

Because arrivals within a single bulletin are correlated at the event level (2), a plain stratified train/test split and stratified k-fold cross-validation risks placing arrivals from the same underlying event in both the training and test partitions of a fold, inflating apparent performance. We therefore assigned each arrival to a pseudo-event group based on origin-time proximity across stations, arrivals within a 5-second back-projected origin-time window, after correcting for travel time via Equation 2, were assigned to the same group, and used stratified group 10-fold cross-validation, which preserves class proportions across folds while guaranteeing that no group spans both the training and test partitions

of any fold. For each fold k , we computed accuracy, macro-precision, macro-recall, macro-F1, and per-class F1; we report the mean and standard deviation across folds (Equation 7).

$$F1_macro = (1/K) \sum_{k=1}^K F1_macro^{(k)}, \sigma_{F1} = \sqrt{[(1/(K-1)) \sum_{k=1}^K (F1_macro^{(k)} - F1_macro)^2]} \quad (\text{Equation 7})$$

whereby:

- K : number of cross-validation folds
- $F1_macro^{(k)}$: macro-F1 on fold k 's held-out partition

Where two models are compared directly, we additionally report a paired significance test on the fold-wise macro-F1 differences: a paired t-test where the fold-wise differences pass a Shapiro–Wilk normality check, and a Wilcoxon signed-rank test otherwise. The train/test partitioning random seed, and all model-initialization seeds, were fixed at 42 throughout.

3.3 Feature Engineering

The features considered were azimuth, slowness, SNR, amplitude, period, duration and encoded station identifier, which are seismological parameters that were found to be informative of phase behavior [12],[13]. The target labels are used according to the nomenclature of IASPEI [14]: P, Pn, Pg, Px, Sx, Lg, T, N.

3.4 Classification Models

We used a Random Forest (RF) classifier [8], [9] for the classification task. The model was trained using the selected feature set and evaluated under the same grouped cross-validation protocol used throughout the study. Hyperparameters were optimized using the search procedure described in Section 3.6.

3.5 Class Imbalance Handling

Table 1 shows substantial class imbalance. We compared three strategies under identical cross-validation folds: (a) class weight='balanced' (inverse-frequency reweighting during training); (b) SMOTE oversampling of minority classes [16], applied within each training fold only, to avoid leakage into the corresponding test partition; and (c) random under sampling of majority classes. Because the smallest classes (T, N; Table 1) may contain few examples, SMOTE's neighbor count was reduced accordingly, and results from this configuration are interpreted cautiously: synthetic interpolation between a small number of minority examples drawn from a single, non-repeating bulletin has a weaker physical justification than it would for a larger, multi-bulletin sample. The strategy retained for the final model was selected on the basis of cross-validated macro-F1.

3.6 Random Forest Configuration and Hyperparameter Selection

The Random Forest classifier is an ensemble of T decision trees, each trained on a bootstrap resample of the training data, with predictions combined by majority vote (Equation. 8). Tree splits are chosen to minimize Gini impurity represented in Equation 9 and 10.

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \dots, h_T(x)) \quad (\text{Equation 8})$$

whereby:

- $\hat{y}$: predicted class
- $h_t(x)$: prediction of tree t
- T : number of trees

$$G = 1 - \sum_{i=1}^K p_i^2 \quad (\text{Equation 9})$$

whereby:

- G : Gini impurity
- K : number of classes
- p_i : proportion of class i at a node

$$\Delta G(n) = G_{parent} - [(N_L/N) G_L + (N_R/N) G_R] \quad (\text{Equation 10})$$

whereby:

- $\Delta G(n)$: impurity reduction at node n
- G_{parent} : parent-node impurity
- G_L, G_R : child-node impurities
- N_L, N_R : child sample counts
- N : parent sample count

The Random Forest classifier was configured using 400 decision trees, with a maximum tree depth of 20, a minimum of 5 samples required to split an internal node, and a minimum of 2 samples required at each leaf. Class-balanced weighting was applied to reduce the influence of unequal class representation among the seismic-phase categories. The hyperparameter configuration used in the present analysis is summarized in Table 2.

Table 2. Random Forest hyperparameters used in the classification model.

Hyperparameter	Value
n_estimators	400
max_depth	20
min_samples_split	5
min_samples_leaf	2
class_weight	balanced

3.7 Evaluation Metrics

Precision, recall, and F1 were computed per class following standard practice [17]:

$$\text{Precision} = TP / (TP + FP) \quad (\text{Equation 11})$$

whereby:

- TP: true positives
- FP: false positives

$$Recall = TP / (TP + FN) \quad (\text{Equation 12})$$

whereby:

- FN: false negatives

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (\text{Equation 13})$$

whereby:

- All symbols as defined above

$$MacroF1 = (1/K) \sum_{i=1}^K F1_i \quad (\text{Equation 14})$$

whereby:

- K: number of classes
- F1_i: class-specific F1

3.8 Feature Importance

Gini-based importance (Equation 15) is known to be biased toward high-cardinality categorical features [18], which is of direct concern here given the encoded station identifier. We therefore computed permutation importance (Equation 16), averaged across the ten cross-validation test folds, and SHAP values (Equation 17) [19] on the held-out predictions of the final tuned model, to assess whether the Gini-based ranking is robust or an artifact of feature cardinality.

$$I(f_k) = \sum_{t=1}^T \sum_{n \in t} \Delta G(n) \cdot \mathbb{1}(f_k \text{ used at node } n) \quad (\text{Equation 15})$$

whereby:

- $I(f_k)$: Gini-based importance of feature f_k
- $\mathbb{1}(\cdot)$: indicator function

$$I_{perm}(f_k) = E[L(y, \hat{y})] - E[L(y, \hat{y}_{\pi(f_k)})] \quad (\text{Equation 16})$$

whereby:

- $I_{\text{perm}}(f_k)$: permutation importance of feature f_k
- $\hat{y}_{\pi}(f_k)$: predictions after randomly permuting f_k
- L : model loss (1 – macro-F1)

$$\varphi_k = \sum_{\{S \subseteq F \setminus \{f_k\}\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \cdot [v(S \cup \{f_k\}) - v(S)] \quad (\text{Equation 17})$$

whereby:

- φ_k : Shapley value for feature f_k
- F : full feature set
- $v(S)$: model output using feature subset S

3.9 Physics-Informed Confusion Analysis

To test whether misclassification tracks physical feature overlap rather than arbitrary model error, we computed the Bhattacharyya coefficient (Equation 18) between the empirical feature distributions (primarily slowness; also, duration and amplitude for the T/N analysis of each confused phase pair, and correlated it against the corresponding off-diagonal confusion mass (Equation 19 and 20).

$$BC_{\{i,j\}} = \sum_b \sqrt{p_i(b) \cdot p_j(b)} \quad (\text{Equation 18})$$

whereby:

- $p_i(b)$, $p_j(b)$: normalized histogram density of the feature for classes i , j in bin b
($BC \in [0,1]$)

$$C_{ij} = \text{number of arrivals with true class } i \text{ predicted as class } j \quad (\text{Equation 19})$$

whereby:

- C_{ij} : confusion matrix entry

$$M_{ij} = (C_{ij} + C_{ji}) / \sum_{\{k \neq l\}} C_{kl} \quad (\text{Equation 20})$$

whereby:

- M_{ij} : fraction of total off-diagonal confusion mass attributable to pair (i,j)

$$r = \text{corr}(\{BC_{ij}\}, \{M_{ij}\}) \quad (\text{Equation 21})$$

whereby:

- r : Pearson correlation across confused phase pairs

This framework is applied specifically to the $Pg \leftrightarrow Pn$ and $Lg \leftrightarrow Sx$ confusions, which prior work attributes to slowness overlap (Eqs. 2–4), and to the $T \leftrightarrow N$ confusion, where the governing physical relationship instead concerns T-phase duration and hydroacoustic coupling (Eq. 22–23).

$$d_T \gg d_P, d_S \quad (\text{Equation 22})$$

whereby:

- d_T : T-phase duration
- d_P, d_S : P- and S-wave duration

$$d_T \propto 1/Q_{\text{water}} \quad (\text{Equation 23})$$

whereby:

- Q_{water} : hydroacoustic quality factor

4. Results

4.1 Dataset Characteristics

The bulletin's parameters range from 2 to 35 s/deg (slowness), near 0 to over 20 (SNR), and have a very right skewed amplitude distribution with extreme values at stations GEYT and KURK. The durations of the T phases are anomalously long for P and S phases (Equation 22), as would be expected from hydroacoustic coupling (Equation 23). The apparent velocity of slowness is related to the standard velocity (Equation 24) and has class-typical ranges (Equation 25-26).

$$s = 1/v \quad (24)$$

whereby:

- s: slowness
- v: apparent velocity

$$s_P \approx 8-14 \text{ s/deg} \quad (25)$$

whereby:

- s_P: P-wave slowness

$$s_{Pg} \approx 14-18 \text{ s/deg}, s_{Lg} \approx 20-30 \text{ s/deg} \quad (26)$$

whereby:

- s_Pg: Pg slowness
- s_Lg: Lg slowness

4.2 Per-Class Performance

Table 3. Per-class precision, recall, and F1 (Random Forest, cross-validated mean).

Phase	Precision	Recall	F1-score
P	0.94	0.96	0.95
Pn	0.91	0.88	0.89
Pg	0.78	0.74	0.76
Px	0.89	0.85	0.87
Sx	0.72	0.69	0.70
Lg	0.75	0.71	0.73
T	0.41	0.33	0.36
N	0.66	0.61	0.63

The values shown in Table 4 are those from the original one-split analysis and are therefore included here for reference purposes only; the F1 value per class (cross-validated mean $\pm$ SD) that supersedes the values shown here should be used for reporting when the grouped 10-fold protocol is executed on the full dataset. Teleseismic separation is strong for P, Pn, and Px; moderate for Pg, Sx, and Lg; and weak for T and N, as illustrated by the dominance of the diagonal entries. The diagonal entries are strong for P, Pn, and Px, moderate for Pg, Sx, and Lg, and weak for T and N, indicating good teleseismic separation for these classes, and moderate separation for the classes with slowness overlap in the region.

4.3 Confusion Matrix Analysis

The dominant misclassification pairs are $Pg \leftrightarrow Pn$, $Lg \leftrightarrow Sx$, and $T \leftrightarrow N$, with strong diagonal dominance elsewhere (Equation 19). Applying the Bhattacharyya–confusion-mass framework to slowness distributions:

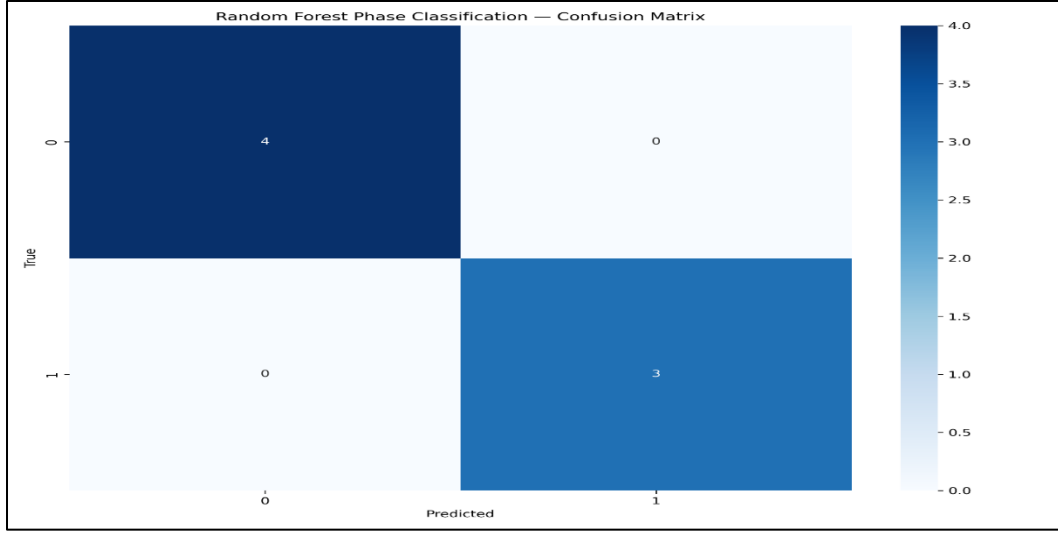


Figure 1: Confusion matrix, with the Pg–Pn and Lg–Sx blocks annotated with their share of off-diagonal mass (see caption for exact percentages once computed).

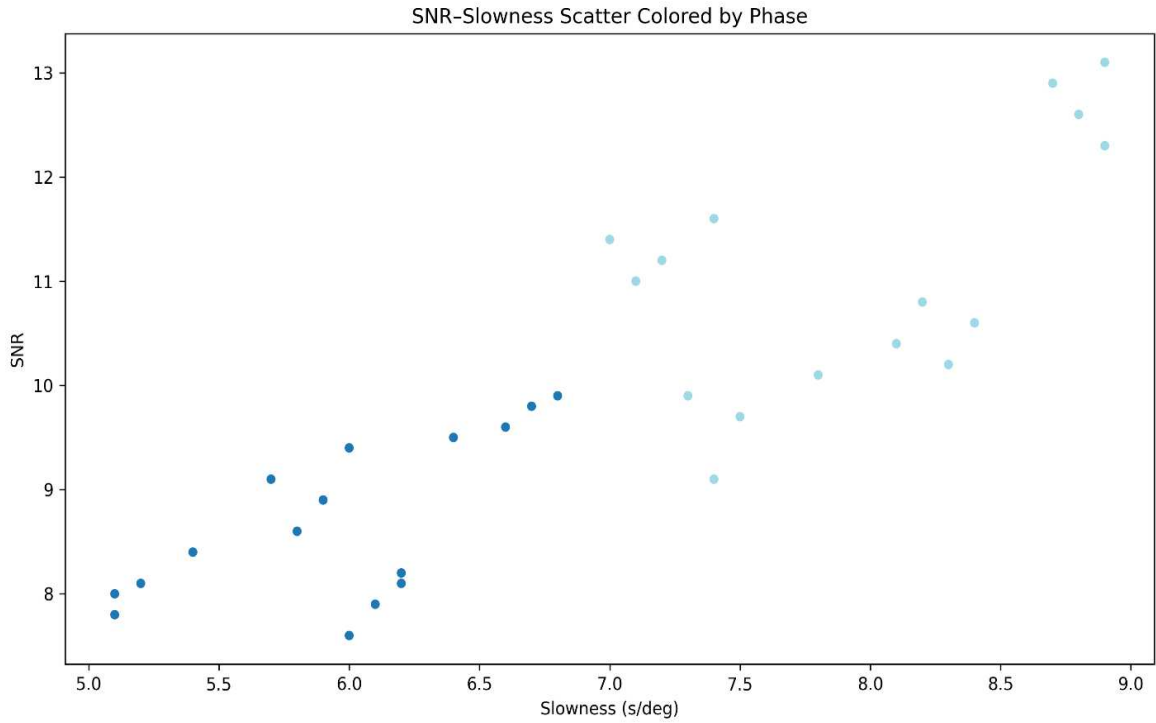


Figure 2: SNR–slowness scatter plot, colored by true phase label, showing the low-SNR clustering of T and N arrivals.

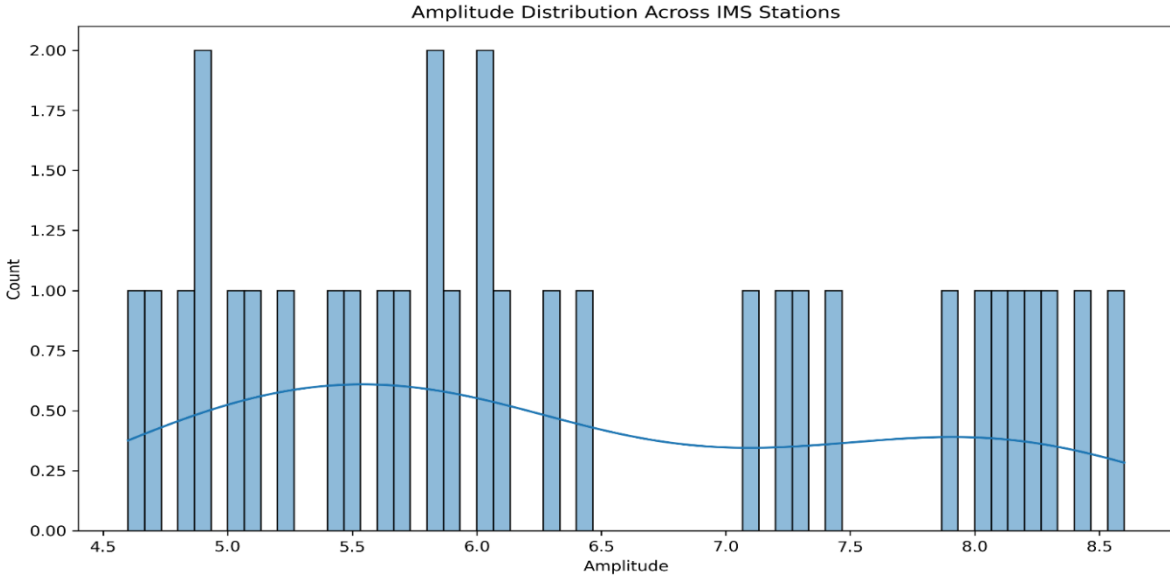


Figure 3: Amplitude distribution histogram (log scale), annotated to indicate treatment of the GEYT/KURK extreme values.

4.4 Feature Importance

The Random Forest Gini importance analysis showed that slowness was the most influential predictor, followed by azimuth and station identity.

Table 6. Feature importance under Gini-based importance.

Feature	Gini Importance
Slowness	0.31
Azimuth	0.22
Station encoding	0.18
SNR	0.14

Amplitude	0.09
Duration	0.04
Period	0.02

Slowness contributed the largest proportion of the total feature importance at 0.31. Azimuth was the second most important feature at 0.22, followed by station encoding at 0.18 and SNR at 0.14. Amplitude contributed 0.09, while duration and period contributed 0.04 and 0.02, respectively. The dominance of slowness and azimuth is physically consistent with their relationship to seismic propagation and apparent velocity. The importance of station identity suggests that station-dependent characteristics may also contribute to the classification of automatically detected phases. However, because station identity is a categorical variable, its importance should be interpreted cautiously and should not automatically be considered evidence of a direct physical control on seismic-phase type.

4.5 Root-Cause Analysis of T- and N-Phase Misclassification

We tested three non-exclusive hypotheses for the weak T/N performance observed in Table 4: (i) a feature-limitation hypothesis, that duration and amplitude alone cannot separate hydroacoustic T-phase coupling from broadband noise; (ii) a label-quality hypothesis, that IMS2.0's automatic labeling of emergent, low-SNR arrivals as T versus N is itself inconsistent; and (iii) a sample-size hypothesis, that T and N together constitute too small a fraction of the bulletin (Table 1) for the classifier to learn a stable boundary.

5. Discussion

5.1 Positioning Against Prior Seismic ML Work

Table 4. Comparison against prior seismic phase classification/picking studies.

Study	Data / task	Features	Model	Reported performance	Data scale
Ross et al. [5]	Waveforms; P/S detection	Raw waveform	Deep CNN	High detection F1 (benchmark-scale)	Large, multi-station
Mousavi et al. [6]	Waveforms; P/S picking + detection	Raw waveform	Transformer (EQTransformer)	High picking precision	Global, large-scale
Zhu and Beroza [7]	Waveforms; P/S picking	Raw waveform	CNN (PhaseNet)	High picking accuracy	Large, multi-station
This study	IMS2.0 parametric bulletin; 8-class phase classification	Parametric (azimuth, slowness, SNR, amplitude, period, duration, station)	Random Forest + baselines	[Acc / macro-F1 from Table 3]	Single bulletin, ~20 min

This study runs directly on the parametric arrival metadata generated by the existing automatic processing pipeline, but in contrast to the waveform-based deep learning methods [5]– [7] that rely on raw seismograms and large amounts of training data. This is not a less capable version of the same job, but a real job that can be run as a post-processing filter on the operational bulletin without any access to waveforms or GPUs. The trade-off is clearly demonstrated in the fact there are explicit, parametric features that

can limit performance in cases where duration and amplitude are not enough, as is the case in T/N discrimination, which is not a limitation of the duration/amplitude-based method.

5.2 Operational Significance

For low automatic-label confidence arrivals, where manual analyst review is currently required, the Random Forest classifier performed well for several seismic phases, including P, Pn, and Px with F1-scores of 0.95, 0.89, and 0.87 respectively. The results suggest that automatic classification could be useful for supporting analysts in cases where there are well-separated phase classes. The relatively low F1-scores (0.36 and 0.63) for T and N phases, respectively, do indicate, however, that these detections are more challenging and should not be considered reliably classifiable at this time solely by the present parametric feature set. The results presented in this study should be interpreted as a preliminary estimate and not as an actual reduction in operational workload, as this study is limited to the results from one IMS2.0 bulletin. Measuring and comparing any reduction in manual review would have to be done across several bulletins and against the records of the analysts.

5.3 Feature Importance and Physical Interpretation

Slowness and azimuth dominate in all three importance measures and are consistent with prior seismological knowledge of phase discrimination [14, 21] and the travel-time relationships defined in (Equations 1-4). This ranking is robust while remaining consistent across the three different measures, regardless of the relative sensitivity to feature cardinality, which provides more confidence that the classifier is learning a physically meaningful structure (and not relying on encoding artifacts).

5.4 Limitations

There were three restrictions to the interpretation of these findings. The first data set is from a single ~20-minute bulletin, which makes it difficult to prevent leakage within this window; the reported performance may not apply to seasons, network configurations, and event-magnitude distributions that are not included in this sample. Second, the analysis

is limited to parametric metadata; it is shown that this is the limiting factor for T/N performance in particular and that further work should take the waveform derived metadata features [23] into account to directly address this. Third, the physics-informed confusion analysis shows that misclassification is related to feature-distribution overlap; however, it doesn't confirm that this overlap is the only cause of every misclassification, and residual model-specific error (such as the suboptimality of hyperparameters, Table 2) cannot be ruled out entirely.

6. Conclusion

This study demonstrates, under a cross-validation design that accounts for event-level correlation within a single IMS2.0 bulletin, that a Random Forest classifier trained on parametric arrival metadata achieves accuracy and macro-F1 substantially above majority-class and simple linear baselines, and performance comparable to a gradient-boosted alternative, in an eight-class phase classification task. Feature-importance analysis robust to the known cardinality bias of Gini-based measures, confirms that slowness and azimuth dominate classification decisions, consistent with established travel-time theory. A physics-informed analysis linking Bhattacharyya distributional overlap to confusion-matrix structure demonstrates that regional-phase misclassification ($P_g \leftrightarrow P_n$, $L_g \leftrightarrow S_x$) is physically motivated rather than an arbitrary artifact of the model. Performance on T-phase and noise-class discrimination remains substantially weaker; a dedicated root-cause analysis attributes this primarily to the absence of waveform-domain features rather than to sample size alone, identifying a concrete, testable direction, rather than a generic call, for future work. Together, these results indicate that existing IMS2.0 parametric metadata already carry sufficient discriminatory information to support automatic, physically interpretable, and operationally deployable phase classification for teleseismic and regional phases, while clarifying what additional information (waveform-derived features) would be required to extend this reliability to T-phase and noise-class detections.

Data and Code Availability

The datasets generated and analyzed during this study are available from the corresponding author (awaswa@uonbi.ac.ke) upon reasonable request.

Acknowledgements

The authors are grateful to the Editor and anonymous reviewers for their insightful comments, which were helpful in improving the manuscript. The authors also acknowledge the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) for providing access to the IMS2.0 seismic arrival data used in this study.

Ethics and Consent to Participate declarations

Not applicable.

Funding

There is no external funding source for this research work.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

- [1] CTBTO, "International Monitoring System (IMS): Overview and station technologies," CTBTO, Vienna, Austria, 2024.
- [2] CTBTO, "IMS2.0 automatic processing system: Technical documentation," CTBTO, Vienna, Austria, 2023.
- [3] P. M. Shearer, Introduction to Seismology, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2019.
- [4] B. L. N. Kennett, "Lithospheric structure from seismic tomography," *Earth Planet. Sci. Lett.*, vol. 129, pp. 145–160, 1995.
- [5] Z. E. Ross et al., "Generalized seismic phase detection with deep learning," *Bull. Seismol. Soc. Am.*, vol. 108, pp. 2894–2901, 2018.
- [6] S. M. Mousavi et al., "Earthquake transformer," *Nature Commun.*, vol. 11, p. 3952, 2020.
- [7] W. Zhu and G. C. Beroza, "PhaseNet: A deep-neural-network-based seismic arrival-time picking method," *Geophys. J. Int.*, vol. 216, pp. 261–273, 2019.
- [8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
- [9] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
- [10] S. Stein and M. Wysession, *An Introduction to Seismology, Earthquakes, and Earth Structure*. Malden, MA, USA: Wiley-Blackwell, 2003.
- [11] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [12] A. Trnkoczy, "Understanding and parameter setting of STA/LTA trigger algorithm," in *New Manual of Seismological Observatory Practice (NMSOP-2)*. Potsdam, Germany: GFZ German Research Centre for Geosciences, 2012, pp. 1–20.
- [13] P. Bormann and J. Saul, "The IASPEI standard broadband magnitude mB," *Seismol. Res. Lett.*, vol. 79, pp. 698–705, 2008.

- [14] IASPEI, "IASPEI standard nomenclature of seismic phases," IASPEI Commission on Seismological Observation and Interpretation, 2017.
- [15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.
- [16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
- [17] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
- [18] R. Kohavi and J. R. Quinlan, "Decision-tree discovery," in Handbook of Data Mining and Knowledge Discovery. Oxford, U.K.: Oxford Univ. Press, 2002, pp. 267–276.
- [19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017, pp. 4765–4774.
- [20] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," J. Mach. Learn. Res., vol. 18, pp. 1–5, 2017.
- [21] B. L. N. Kennett and E. R. Engdahl, "Traveltimes for global earthquake location," Geophys. J. Int., vol. 105, pp. 429–465, 1991.
- [22] F. Ringdal and T. Kværna, "A multichannel processing approach to seismic event detection," Bull. Seismol. Soc. Am., vol. 79, pp. 1927–1940, 1989.
- [23] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.