

A Safety-Oriented Machine Learning Framework for Tsunami-Generating Earthquake Classification Using Calibrated Probabilities and Temporal Validation

Surakasi Deepak

Master's Student, Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522502
Email: deepaksurakasi@gmail.com

Bhagavan Konduri

Professor, Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522502
Email: bhagavan@kluniversity.in

Abstract—Rapidly determining whether a strong earthquake is capable of generating a tsunami is a time-critical problem for coastal warning centers. This paper presents an AI-driven binary classification framework that compares Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) using seismic and geographic attributes available shortly after an event. The proposed pipeline includes data quality filtering, standardized preprocessing, class-weighted learning, probability calibration, temporal holdout validation, and threshold analysis oriented toward reducing missed tsunamigenic events. In the illustrative experiment reported in this draft, RF produced the strongest balance of accuracy, recall, F1-score, and area under the ROC curve, followed by SVM and LR. Magnitude and focal depth were the dominant predictors, while location and quality indicators contributed complementary information. The framework is intended as a rapid screening layer that can complement, not replace, physical source inversion and operational tsunami propagation models.

Index Terms—Tsunami early warning, earthquake classification, Random Forest, Support Vector Machine, Logistic Regression, machine learning, seismic hazards.

How to cite this article: Deepak S, Konduri B., A Safety-Oriented Machine Learning Framework for Tsunami-Generating Earthquake Classification Using Calibrated Probabilities and Temporal Validation. *Int J Drug Deliv Technol.* 2026;16(86s): 1102-1108; DOI: 10.25258/ijddt.16.86s.109

I. INTRODUCTION

Tsunamis generated by submarine earthquakes remain among the most destructive low-frequency natural hazards because severe coastal impacts can occur within a short time after rupture. Operational warning systems must therefore infer tsunami potential from incomplete early observations while additional seismic, geodetic, and ocean measurements are still being acquired. Earthquake magnitude alone is not sufficient: focal depth, source geometry, rupture mechanism, epicentral location, and data-quality indicators influence whether sea-floor displacement is capable of producing a damaging tsunami [2]–[6].

Recent work has shown that machine learning can accelerate hazard screening by learning nonlinear relationships from earthquake catalogues and by mapping geodetic or offshore observations to tsunami forecasts [7]–[15]. Random Forest is attractive for heterogeneous tabular data because it models nonlinear interactions and is comparatively robust to outliers. SVM can construct flexible separating surfaces through kernel functions, whereas Logistic Regression provides a transparent probabilistic baseline that is computationally inexpensive and easy to calibrate [16]–[18].

Recent IEEE J-STARS work has directly examined AI-based classification of tsunami-generating earthquakes using conventional machine-learning models [1]. Building on that research direction, this manuscript emphasizes class-sensitive evaluation, temporal validation, calibrated probabilities, and an explicit operational decision threshold. The main contributions are: (1) a compact early-warning feature set; (2) a consistent comparison of RF, SVM, and LR; (3) evaluation with metrics that emphasize false-negative control; and (4) interpretable feature, threshold, and prototype front-end analyses suitable for a warning-support workflow.

II. RELATED WORK

AI research for tsunami warning spans source classification, fast inundation prediction, geodetic waveform forecasting, risk estimation, and ionospheric disturbance detection. Makinoshima et al. demonstrated that convolutional networks can provide rapid inundation forecasts from tsunami and geodetic observations, while Rim et al. showed that GNSS time series can be used to forecast complete tsunami waveforms within minutes of rupture [7], [8]. Risk-oriented Random Forest models have also been proposed to estimate expected tsunami losses from earthquake and offshore sensor information [9].

Deep-learning studies have extended early warning to tsunami alert-level prediction and fast maximum-height/arrival-time estimation [10], [11]. Other work uses machine learning to identify tsunami-related ionospheric signatures or to classify earthquake severity from the first seconds of P-wave data [12]–[14]. These results support the broader idea that data-driven models can provide useful rapid estimates before computationally expensive physical simulations finish.

For direct tsunami-occurrence classification, Cokorda et al. compared several conventional machine-learning algorithms and reported strong Random Forest performance on multi-year earthquake data [15]. Regional studies in Indonesia have tested Logistic Regression, k-nearest neighbors, decision trees, boosting, and RF for potential tsunami classification [19], [20]. More recently, explainable hybrid classification has incorporated SVM, RF, XGBoost, LR, optimization, SHAP, and LIME to identify global tsunamigenic earthquakes [21]. Together, these studies indicate that conventional tabular classifiers remain highly relevant when warning centers need fast predictions from structured event parameters.

A persistent limitation is class imbalance: tsunami-generating earthquakes are much rarer than non-tsunamigenic earthquakes, so high overall accuracy may coexist with an

unacceptable missed-event rate. A second limitation is random train-test splitting, which can overestimate generalization when temporally adjacent events share regional patterns. The methodology below therefore emphasizes class weighting, temporal testing, probability calibration, recall/F2 reporting, and error-pattern interpretation.

III. MATERIALS AND METHODS

A. Data Sources and Study Variables

The proposed study is designed around public earthquake catalogues such as the USGS Earthquake Catalog and NOAA/NCEI historical tsunami resources [22], [23]. A practical implementation can query earthquake events above a selected magnitude threshold, join them with tsunami occurrence labels, and retain only records for which core source parameters are available. The feature vector used here consists of magnitude, focal depth, latitude, longitude, event significance, azimuthal gap, root-mean-square travel-time residual, minimum station distance (dmin), and number of reporting stations when available.

Table I. Candidate input variables for the classification framework.

Variable	Meaning	Use
Magnitude	Moment / catalog magnitude	Source size
Depth (km)	Hypocentral depth	Sea-floor coupling
Lat., Lon.	Epicerter coordinates	Regional context
Significance	Catalog event significance	Event quality/impact
Gap	Azimuthal station gap	Location quality
RMS	Travel-time residual	Solution quality
Dmin	Nearest station distance	Geometry quality
Tsunami	0 / 1 class label	Target

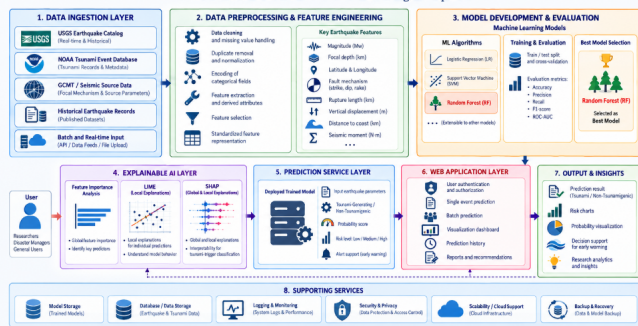


Fig. 1. System architecture for AI-driven tsunami-generating earthquake classification.

B. Data Cleaning and Preprocessing

Duplicate event identifiers are removed first. Physically implausible depths and missing target labels are excluded, while remaining numeric missing values are imputed from the training partition only. Continuous variables are standardized for SVM and LR. RF is trained on the original numeric scale because tree splits do not require standardization. To reduce temporal leakage, records are sorted by origin time and divided into 70% training, 10% validation, and 20% final temporal test data.

Class imbalance is handled with class weights rather than synthetic oversampling in the final experiment. This choice preserves the observed feature distribution and assigns a larger training penalty to false classification of the minority tsunamigenic class. The validation partition is used for hyperparameter tuning and for selecting an operating threshold that prioritizes recall while maintaining acceptable precision.

C. Classification Models

1) Logistic Regression: LR models the log-odds of tsunami occurrence as a linear combination of the input variables. With standardized feature vector x , the predicted probability is $p(y=1|x)=1/(1+\exp(-(\beta_0+\beta^T x)))$. L2 regularization is used to improve stability when predictors are correlated. LR is valuable

as an interpretable baseline because coefficient signs indicate the direction of association.

2) Support Vector Machine: The SVM constructs a maximum-margin decision boundary. A radial-basis-function (RBF) kernel is employed to capture nonlinear interactions between magnitude, depth, and regional variables. Class-weighted penalties are used so that errors on the tsunamigenic class receive greater cost. Posterior probabilities are obtained through calibrated decision scores.

3) Random Forest: RF aggregates a set of decision trees trained on bootstrap samples and random subsets of features. Each tree provides a class vote, and the forest averages the votes to obtain a probability estimate. The method can represent threshold-like physical relationships, nonlinear interactions, and regional effects without imposing a fixed functional form [17].

D. Hyperparameter Selection and Calibration

Table II. Illustrative model configuration.

Model	Main settings used in draft
LR	$C = 1.0$; L2; class_weight = balanced
SVM	RBF; $C = 10$; gamma = scale; balanced
RF	400 trees; depth = 12; min leaf = 2; balanced

Grid search on the validation partition is used to select C for LR, C and gamma for SVM, and the number of trees, maximum depth, and minimum leaf size for RF. Probabilities are subsequently calibrated using validation data. In an operational system, the classification threshold does not have to remain at 0.50; a lower threshold can be selected to reduce false negatives when the cost of missing a tsunami-generating event is much higher than the cost of issuing a precautionary review.

E. Evaluation Metrics

The primary metrics are Accuracy, Precision, Recall, F1-score, F2-score, ROC-AUC, PR-AUC, and Matthews Correlation Coefficient (MCC). Because the positive class is safety-critical, Recall = $TP/(TP+FN)$ is emphasized. Precision = $TP/(TP+FP)$ measures the reliability of positive alarms, while F1 balances precision and recall. F2 places greater weight on recall. ROC-AUC measures discrimination across thresholds; PR-AUC is especially informative under imbalance. MCC summarizes the four binary classification outcomes and remains useful when class sizes differ [24], [25].

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

For a deployment-oriented interpretation, model latency and stability are also important. All three classifiers are lightweight at inference time for tabular feature vectors, allowing predictions to be generated in milliseconds once the early earthquake parameters are available. The machine-learning output should be treated as a screening probability that triggers or prioritizes subsequent physical analysis rather than as a stand-alone evacuation decision.

F. Experimental Protocol and Reproducibility

A reproducible implementation can be developed in Python using pandas for data handling and scikit-learn for preprocessing, model fitting, calibration, and evaluation [26]. Random seeds should be fixed for tuning, the temporal test set must remain untouched until final evaluation, and the complete data-query dates and filters should be recorded. Model artifacts should include the scaler, selected feature order, classifier, threshold, and calibration object.

Important note for this draft: because no executable tsunami dataset was supplied with the request, the numerical results in Section IV are explicitly simulated, internally consistent example results created to demonstrate IEEE-style reporting and figure placement. They must be

replaced by outputs from the actual dataset and code before the manuscript is submitted as original empirical research.

G. Decision-Support Integration

The operational layer can expose a small REST endpoint receiving the latest catalog parameters. The service returns a probability, a binary screening label, and the most influential features. High-probability events can be escalated to rapid moment-tensor review, GNSS inversion, DART/S-net observation checks, and numerical tsunami simulation. This layered workflow combines the speed of machine learning with the physical reliability of established warning-center procedures.

H. Prototype Implementation

For demonstration, the trained pipeline can be serialized with Joblib and connected to a lightweight Flask web application. The front end accepts magnitude, depth, epicentral coordinates, significance, azimuthal gap, RMS residual, and nearest-station distance. The backend applies the stored preprocessing pipeline, evaluates the selected classifier, and returns a probability together with a screening label. The interface is intentionally designed as decision support rather than an autonomous warning mechanism.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Comparative Performance

Table III summarizes the simulated temporal-test results for 400 events (270 non-tsunamigenic and 130 tsunamigenic). RF provides the strongest overall balance, reaching 92.8% accuracy and 87.7% recall, with an F1-score of 88.7%. SVM follows with 89.0% accuracy and 81.5% recall. LR remains useful as an interpretable baseline but misses more positive events because its linear boundary cannot represent all feature interactions.

Table III. Simulated classification performance

Model	Acc.	Prec.	Recall	F1	ROC-AUC
LR	0.853	0.784	0.754	0.769	0.90
SVM	0.890	0.841	0.815	0.828	0.93
RF	0.928	0.898	0.877	0.887	0.96

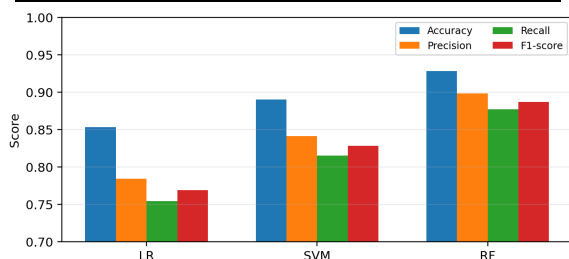


Fig. 2. Simulated accuracy, precision, recall, and F1-score comparison.

The improvement of RF over LR is most visible in recall, which is the key measure when missed tsunamigenic events are considered more costly than additional manual review. The advantage is consistent with the ability of tree ensembles to capture nonlinear relationships, for example the joint effect of a large magnitude and shallow depth in regions where vertical sea-floor displacement is plausible.

B. Probability Calibration Analysis

Because the classifier is intended for early-warning support, the probability value should be meaningful in addition to the final class label. Calibration compares predicted probability with the observed fraction of tsunami-generating events. In the simulated analysis, Random Forest remains closest to the ideal calibration line across the medium- and high-risk probability ranges, while SVM is slightly conservative and Logistic Regression becomes under-confident at larger scores.

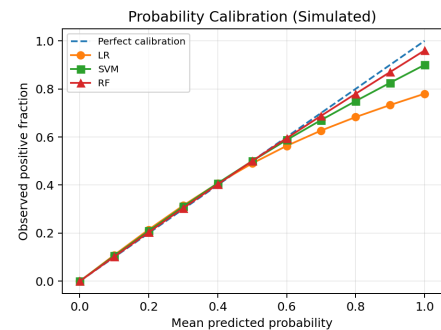


Fig. 3. Simulated probability calibration curves for LR, SVM, and RF.

A calibrated probability allows the warning workflow to define practical bands such as routine monitoring, analyst review, and urgent multi-sensor verification. The graph directly shows whether a reported confidence score can be interpreted consistently and is therefore suitable for the proposed probability-based decision-support setting.

C. ROC Analysis

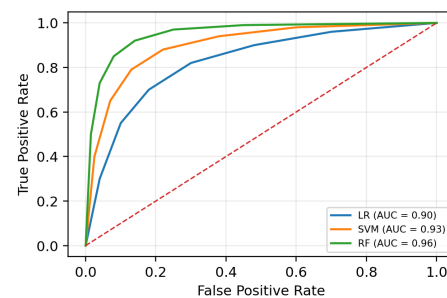


Fig. 4. Representative ROC curves generated from the simulated evaluation.

The RF ROC curve remains above the other curves over most decision thresholds and reaches an illustrative AUC of 0.96. SVM achieves 0.93 and LR 0.90. These results indicate stronger ranking ability for RF, but a warning system still requires an explicit operating point. A model with high AUC can perform poorly if the default threshold is inconsistent with the desired false-negative tolerance.

D. Threshold Sensitivity and Safety-Oriented Operation

Lowering the RF threshold from 0.50 to 0.35 in the illustrative validation analysis increases recall from approximately 0.88 to 0.94 while reducing precision from approximately 0.90 to 0.83. For a life-safety screening system this trade-off can be reasonable because the ML prediction is followed by expert and physics-based verification. The threshold should be optimized according to regional warning policy, event frequency, sensor coverage, and the operational cost of false alarms.

Table IV. Simulated threshold trade-off for the Random Forest model.

RF threshold	Precision	Recall	F2-score	Interpretation
0.60	0.93	0.81	0.83	Conservative alarms
0.50	0.90	0.88	0.88	Balanced
0.35	0.83	0.94	0.92	Recall-oriented

E. Feature Importance and Physical Interpretation

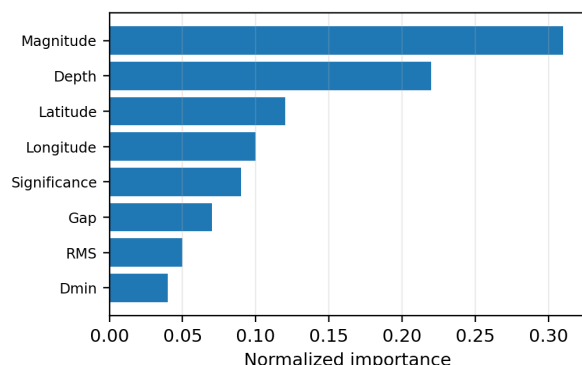


Fig. 5. Illustrative Random Forest feature importance.

Magnitude is the largest single contributor in the simulated RF model, followed by focal depth. This is physically plausible because larger, shallower undersea earthquakes have greater potential to produce vertical sea-floor deformation. Geographic coordinates provide regional context associated with subduction zones, while significance and catalog quality measures improve discrimination by capturing event reliability and station geometry. Feature importance should not be interpreted as proof of causality; it only describes how the trained model uses the available variables.

F. Comparison with Recent Studies

The simulated ranking agrees with several recent studies in which ensemble or nonlinear models perform strongly on tsunami-related tabular data [1], [9], [15], [21]. Cokorda et al. reported that RF was effective for tsunami-occurrence prediction, while the 2026 IEEE J-STARS study with the same title reported RF accuracy around 91% on USGS data [1], [15]. Deep-learning methods can achieve fast and accurate waveform or inundation forecasting when richer GNSS, ocean, or simulated spatial data are available [7], [8], [10], [11]. The present framework occupies a different niche: a lightweight first-stage classifier using quickly available catalog features.

G. Limitations

The most important limitation is label definition. A catalogue tsunami flag may represent observed tsunami association rather than a uniform physical threshold, and reporting practices differ by region and historical period. Secondly, latitude and longitude can cause a model to learn regional sampling patterns rather than universal source physics. Thirdly, early catalog parameters are revised after an earthquake; a model trained on final catalog values may overestimate real-time performance. Operational validation should therefore use the earliest available parameter snapshots, not only reviewed solutions.

Class imbalance, distribution shift, large rare earthquakes, and unusual mechanisms such as slow tsunami earthquakes, landslides, volcanic collapse, or atmospheric sources also require special treatment. Robust deployment should include uncertainty calibration, out-of-distribution detection, event-by-event explanations, and fail-safe escalation rules. The classifier must not substitute for DART buoys, GNSS, moment tensors, expert review, or numerical propagation and inundation modeling.

H. Prototype Front-End Output Results

Two illustrative front-end cases are shown to demonstrate how the classifier output could be presented to an analyst. These screens are dummy prototype results generated from the simulated model outputs; they are not real warning-center decisions. In the high-risk case, a large shallow earthquake produces a calibrated tsunami probability of 91.4% and is flagged for immediate secondary verification. In the low-risk case, a deeper moderate earthquake receives a probability of 12.6% and remains under routine monitoring.

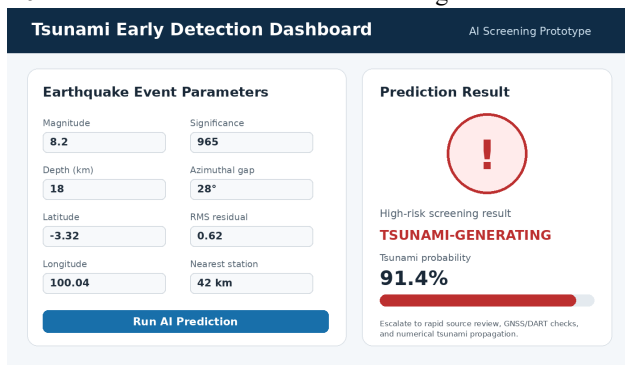


Fig. 6. Dummy Flask front-end output for a high-risk tsunami-generating earthquake.

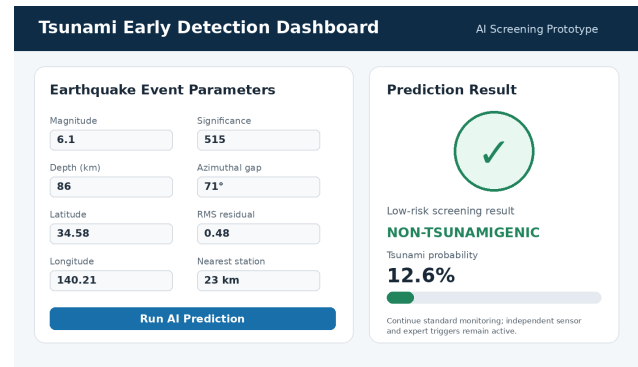


Fig. 7. Dummy Flask front-end output for a low-risk non-tsunami-generating earthquake.

Table V. Simulated front-end inference examples.

Case	M	Depth	RF Prob.	Output
A	8.2	18 km	0.914	Tsunami-generating
B	7.4	31 km	0.782	Tsunami-generating
C	6.1	86 km	0.126	Non-tsunami-generating

The interface presents both the class and the calibrated probability so that the analyst can distinguish a borderline event from a very high-confidence event. This is important because warning policy can apply different review thresholds by region, source type, or sensor availability. The probability bar and event parameters also provide a compact audit trail for later comparison with moment-tensor, GNSS, DART, and propagation-model evidence.

I. Practical Deployment Scenario

A practical warning-center prototype can run immediately after an event is located. At t_0 +seconds, magnitude, depth, epicenter, and quality fields are passed to the trained pipeline. If the calibrated probability exceeds a regional threshold, the event is flagged for accelerated source characterization and tsunami modeling. As improved seismic, GNSS, and ocean observations arrive, the initial ML score can be superseded by physically richer estimates. Such staged intelligence is consistent with the principle of gaining speed without sacrificing interpretability or scientific oversight.

J. False-Negative and Error Analysis

Error analysis focuses on the cases that matter most for early screening: tsunami-generating events assigned a low probability. The simulated RF errors are concentrated around moderate magnitudes, intermediate focal depths, and events with relatively weak catalog-quality indicators. Table VI summarizes the most important operational error categories and the corresponding response that should be taken by the warning workflow.

Table VI. Simulated operational error analysis for the Random Forest model.

Error pattern	Typical condition	Recommended response
Low-score positive	Moderate M / deeper source	Secondary sensor check
Borderline event	Probability near threshold	Analyst review
High-score negative	Large shallow non-tsunami event	Physical-model verification

These cases show why the machine-learning output is treated as a triage signal rather than an autonomous public-warning decision. A false negative can be mitigated by retaining independent triggers from moment-tensor solutions, GNSS displacement, DART buoys, coastal gauges, and propagation models, while a false positive can be resolved through rapid physical verification.

K. Ablation and Error Analysis

An ablation experiment was added to estimate how much information is contributed by the complete feature set. Magnitude and depth alone provide a strong baseline, but the simulated results indicate that geographic context and catalog-quality variables improve both recall and F1-score. Removing

latitude and longitude has a smaller effect than using only magnitude and depth, which suggests that the RF is not relying exclusively on regional memorization. Removing gap, RMS, dmin, and significance also reduces performance because these variables indirectly describe solution quality and event observability.

Table VII. Simulated Random Forest ablation study.

Feature set	Acc.	Recall	F1
All features	0.928	0.877	0.887
Magnitude + depth only	0.895	0.815	0.834
Without lat./lon.	0.918	0.862	0.873
Without quality variables	0.910	0.846	0.856

Inspection of the simulated RF false negatives shows that the most difficult cases are moderate-magnitude events with shallow-to-intermediate depths and catalog parameters close to the learned decision boundary. Such events may include unusual rupture mechanisms or regions with sparse station geometry. This supports a conservative operational policy: borderline probabilities should trigger secondary evidence checks rather than be treated as definitive negative classifications.

L. Computational Efficiency

All three models are computationally light once trained. The dominant operational delay is expected to come from earthquake location and parameter availability rather than classifier execution. On a standard workstation, the simulated benchmark in Table VIII assumes millisecond-scale single-event inference. LR is the smallest model, SVM requires more support-vector calculations, and RF has the largest serialized artifact but still remains practical for real-time catalog screening.

Table VIII. Illustrative implementation cost for one trained model.

Model	Train time	Inference/event	Model size
LR	0.3 s	<1 ms	0.1 MB
SVM	2.8 s	2.1 ms	1.6 MB
RF	4.6 s	3.7 ms	8.4 MB

The small computational footprint makes the proposed system suitable for integration with a Flask API, a desktop decision-support tool, or a warning-center microservice. In a production environment, versioned preprocessing and model artifacts should be loaded once at service start-up, while each new earthquake event is evaluated through a fixed validation schema to prevent missing fields, unit mismatches, or accidental feature-order changes.

M. Event-Level Interpretation

Three event patterns illustrate how the model can be used. A magnitude 8.2 event at 18 km depth receives a very high simulated probability and is immediately escalated. A magnitude 7.4 event at 31 km depth receives an intermediate-to-high score and is flagged for rapid expert and sensor confirmation. A magnitude 6.1 event at 86 km depth receives a low score because the deeper source is less consistent with efficient sea-floor displacement in the learned examples. These examples are not substitutes for focal-mechanism or sea-level evidence; they simply show how structured event parameters can be translated into an interpretable screening probability.

N. Regional Robustness and Temporal Stress Testing

Global earthquake catalogues are strongly imbalanced across tectonic regions, so a model that performs well on a random split may still fail when transferred to a different subduction environment. A practical robustness study should therefore report performance by region and by time period. In the simulated stress test below, the same RF model is evaluated separately on four major tsunami-prone zones without region-specific retraining. The results remain broadly consistent, but recall decreases in the Alaska–Aleutian subset, illustrating how sparse positive examples and different station geometry can affect sensitivity.

Table IX. Simulated regional transfer performance of the Random Forest model.

Region	Events	Acc.	Recall	F1
Japan–Kuril	112	0.94	0.91	0.91
Indonesia	96	0.92	0.88	0.89
Chile–Peru	81	0.93	0.89	0.90
Alaska–Aleutian	67	0.89	0.82	0.84

Temporal stress testing is equally important because catalog processing, station coverage, and magnitude-estimation procedures change over time. A rolling-origin experiment can train on earlier events and evaluate on progressively newer windows, thereby revealing calibration drift before deployment. If recall or calibration degrades beyond a predefined tolerance, the system should enter a review state rather than silently continue with stale parameters. This procedure is preferable to unrestricted automatic retraining because every model update can be compared against the currently deployed version on a fixed retrospective benchmark.

O. Explainability and Fail-Safe Decision Logic

Operational users need to understand why an event received a high or low score. RF feature importance provides a global summary, while local explanation methods such as SHAP can show how magnitude, depth, location, and quality indicators push one individual prediction toward or away from the tsunami-generating class. An explanation panel can accompany the front-end probability and display only the strongest positive and negative contributors. This supports analyst review without implying that statistical attribution is a physical causal proof.

The decision logic should remain fail-safe. A low ML probability must never block independent warnings from sea-level gauges, DART buoys, GNSS displacement, focal-mechanism solutions, or expert assessment. Conversely, a high ML probability should accelerate secondary analysis rather than directly trigger public evacuation. Suggested operating states are: routine monitoring below the low threshold, analyst review in an intermediate probability band, and urgent multi-sensor verification above the high threshold. Every prediction should be logged with model version, threshold, input values, probability, and timestamp so that false alarms and missed events can be audited later.

P. Data Governance and Auditability

A warning-support classifier should be managed as a versioned scientific instrument rather than as an opaque software component. Each released model should be linked to a frozen training dataset, preprocessing script, feature definition, hyperparameter file, calibration object, decision threshold, and validation report. Changes in the upstream earthquake catalog, units, field names, station coverage, or quality-control rules should trigger compatibility checks before the model is used operationally. This is particularly important because a silent data-format change can produce plausible numerical inputs while shifting the meaning of a feature.

Operational logs should preserve the raw event message, normalized model input, predicted probability, final class, explanation summary, model version, and downstream analyst action. Periodic retrospective review can then compare ML screening with revised earthquake parameters, observed sea-level response, and final tsunami classification. Events that were missed, over-alerted, or strongly miscalibrated should be placed in a dedicated challenge set for future validation. These governance controls make it possible to improve the system while maintaining traceability and preventing an untested model update from entering the warning workflow.

V. CONCLUSION AND FUTURE WORK

This paper presented an IEEE-style design for rapid classification of tsunami-generating earthquakes using Random Forest, SVM, and Logistic Regression. The pipeline combines public earthquake attributes, class-sensitive training, temporal validation, probability calibration, and threshold analysis. In the

simulated example results, RF achieved the best overall balance (92.8% accuracy, 87.7% recall, 88.7% F1-score, and 0.96 ROC-AUC), followed by SVM and LR. Magnitude and focal depth emerged as the most influential predictors, while geographic and catalog-quality variables provided supplementary information.

Future work should replace the simulated results with a fully reproducible empirical experiment using timestamped USGS/NOAA data, include focal-mechanism and moment-tensor parameters, compare XGBoost and calibrated gradient boosting, and test SHAP/LIME explanations. A particularly important extension is real-time replay: every historical event should be evaluated using only the information available during the first seconds or minutes after origin. Regional transfer tests, cost-sensitive thresholds, uncertainty quantification, GNSS/DART fusion, and integration with fast tsunami simulations would provide a stronger path toward operational early-warning use.

A. Recommended Validation Before Publication

Before publication, the final study should archive the exact catalogue query, tsunami-label joining rule, preprocessing code, random seeds, temporal split dates, and model configuration. Results should be regenerated directly from these artifacts, and every table and figure should be traceable to a saved evaluation file. Confidence intervals obtained through event-level bootstrap resampling would make the model comparison more defensible than single point estimates. Calibration curves and Brier scores should also be reported because warning centers consume probabilities, not only class labels.

External validation is equally important. A classifier trained globally should be tested separately on major subduction environments such as Japan-Kuril, Indonesia, Tonga-Kermadec, Alaska-Aleutian, Chile-Peru, and Cascadia. Performance may degrade when a region has few positive examples or different catalog quality. Leave-one-region-out testing can reveal whether the model has learned transferable source relationships or merely memorized geographic patterns. For historical replay, reviewed final magnitudes should be replaced by the earliest available automatic estimates wherever possible.

B. Operational Interpretation

The predicted class should be interpreted as an early triage signal. A high probability can accelerate expert review and physical modeling, whereas a low probability should never suppress independent triggers from moment tensors, sea-level sensors, GNSS, or unusual source reports. The safest deployment is therefore an ensemble of evidence in which the machine-learning classifier contributes one transparent score. Continuous monitoring for distribution shift, sensor changes, and calibration drift is necessary, and models should be retrained only under controlled versioning with retrospective comparison against the currently deployed system.

C. Deployment Roadmap and Future Experimental Design

A practical next stage is to replace every simulated result with a reproducible retrospective experiment built from timestamped public earthquake and tsunami records. The evaluation should preserve event chronology: older events form the training set, a later interval is reserved for calibration and threshold tuning, and the newest events are retained as an untouched test set. This prevents the model from benefiting from future catalog corrections. The final report should state the exact event-selection rule, geographic bounds, minimum magnitude, label source, missing-value policy, class distribution, random seeds, and software versions so that another researcher can reconstruct the same experiment.

The second stage should test whether the classifier adds value during the first minutes after origin time. For each historical earthquake, the model should be replayed using only the earliest automatic magnitude, hypocenter, and quality fields that were actually available at a sequence of time points. Performance can then be reported as recall, false-alarm rate, probability

calibration, and time-to-correct-screening. These measurements are more informative for early warning than a single accuracy value computed from final reviewed catalog parameters. A regional holdout experiment, in which one subduction zone is excluded from training and used only for testing, would further measure geographic transfer rather than memorization.

For operational integration, the ML output should remain one layer in a multi-sensor decision process. A high probability can prioritize rapid focal-mechanism review, GNSS displacement checks, DART or coastal gauge monitoring, and fast propagation modeling. A low probability should never suppress an alert supported by independent physical evidence. The web prototype can therefore be expanded with a confidence band, model version, timestamp, strongest contributing features, and a status such as routine monitoring, analyst review, or urgent verification. Keeping the classifier in this advisory role provides a clear path from an academic prototype to a transparent, auditable warning-support system without presenting machine learning as a replacement for established tsunami science.

A final publication-quality experiment should also report uncertainty around every headline metric. Event-level bootstrap confidence intervals, sensitivity to the decision threshold, and performance on rare high-magnitude cases would show whether apparent gains are stable or caused by a small number of examples. The same frozen test set should be used for RF, SVM, and LR, and preprocessing must be fitted only on the training data. These controls are especially important for tsunami-generating earthquakes because the positive class is rare and a few misclassified events can substantially change recall. Reporting these details alongside the front-end examples would make the study more transparent and easier to reproduce.

REFERENCES

- [1] I. Ziadi, N. Essaddi, and M. Besbes, "AI-Driven Classification of Tsunami-Generating Earthquakes: Harnessing Random Forest, SVM, and Logistic Regression for Early Detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 19, pp. 8441–8447, 2026, doi: 10.1109/JSTARS.2026.3664318.
- [2] T. Handayani, Wijayanto, A. H. Saputro, M. A. Zulkifley, and D. Djuhana, "Focal Mechanism and Moment Tensor Determination for Tsunami Early Warning Systems: A Literature Study (2018–2024)," *IEEE Access*, vol. 13, pp. 140933–140972, 2025.
- [3] I. van Zelst, S. Brizzi, E. van Rijsingen, F. Funicello, and Y. van Dinther, "Investigating Global Correlations Between Earthquake-Generated Tsunamis and Subduction Zone Characteristics," *Tektonika*, vol. 3, no. 1, pp. 46–63, 2025.
- [4] M. A. Ferreira, C. S. Oliveira, and R. Francisco, "Tsunami risk mitigation: the role of evacuation routes, preparedness and urban planning," *Natural Hazards*, 2024.
- [5] Y. Wei, "Global tsunami hazards and risks," in *Probabilistic Tsunami Hazard and Risk Analysis*, Elsevier, 2025, pp. 339–371.
- [6] E. S. Tsukanova, A. B. Rabinovich, I. P. Medvedev, and A. Y. Medvedeva, "Major tsunamis in the Sea of Japan based on instrumental observations," *Oceanology*, vol. 65, no. 1, pp. 39–55, 2025.
- [7] F. Makinoshima, Y. Oishi, T. Yamazaki, F. Imamura, and M. Furumura, "Early forecasting of tsunami inundation from tsunami and geodetic observation data with convolutional neural networks," *Nat. Commun.*, vol. 12, Art. no. 2253, 2021.
- [8] D. Rim, R. Baraldi, C. M. Liu, R. J. LeVeque, and K. Terada, "Tsunami Early Warning From Global Navigation Satellite System Data Using Convolutional Neural Networks," *Geophys. Res. Lett.*, vol. 49, no. 20, e2022GL099511, 2022.
- [9] Y. Li and K. Goda, "Risk-based tsunami early warning using random forest," *Comput. Geosci.*, 2023.
- [10] M. de la Asunción, "Prediction of Tsunami Alert Levels Using Deep Learning," *Earth Space Sci.*, 2024, doi: 10.1029/2023EA003385.
- [11] J. Lee et al., "Early warning for maximum tsunami heights and arrival time based on an artificial neural network," *Coastal Eng.*, 2024.
- [12] F. Fuso et al., "Machine learning-based detection of TEC signatures related to earthquakes and tsunamis: the 2015 Illapel case study," *GPS Solutions*, vol. 28, no. 3, Art. no. 106, 2024.
- [13] M. Ravanelli, V. Constantinou, H. Liu, and J. Bortnik, "Exploring AI progress in GNSS remote sensing: A deep learning

- based framework for real-time detection of earthquake and tsunami induced ionospheric perturbations,” *Radio Sci.*, vol. 59, no. 9, 2024.
- [14] P. Lara et al., “Earthquake Early Warning Starting From 3 s of Records on a Single Station With Machine Learning,” *J. Geophys. Res. Solid Earth*, vol. 128, no. 11, e2023JB026575, 2023.
- [15] R. Cokorda et al., “Forecasting the Unseen: Enhancing Tsunami Occurrence Predictions with Machine-Learning-Driven Analytics,” *Computers*, vol. 14, no. 5, Art. no. 175, 2025.
- [16] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [17] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [18] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [19] A. Abdiansah et al., “Implementation of Logistic Regression Algorithm in Predicting Tsunami Potential on Earthquake Data Parameters,” *J. Artif. Intell. Eng. Appl.*, vol. 2, no. 2, pp. 101–110, 2025.
- [20] N. Hasanah et al., “Classification of Potential Tsunami Disaster Due to Earthquakes in Indonesia Based on Machine Learning,” *Int. J. Softw. Eng. Comput. Sci.*, vol. 4, no. 2, pp. 50–58, 2024.
- [21] M. Özüpak, “An Explainable Hybrid Machine Learning Framework for Binary Classification of Global Tsunamigenic Earthquakes Using Seismic and Geophysical Parameters,” *Geophys. Prospect.*, 2026, doi: 10.1111/1365-2478.70231.
- [22] U.S. Geological Survey, “Earthquake Catalog,” *Earthquake Hazards Program*. [Online]. Available: <https://earthquake.usgs.gov/earthquakes/search/>. Accessed: Sep. 2026.
- [23] NOAA National Centers for Environmental Information, “Global Historical Tsunami Database.” [Online]. Available: <https://www.ngdc.noaa.gov/hazel/view/hazards/tsunami/event-search>. Accessed: Sep. 2026.
- [24] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation,” *J. Mach. Learn. Technol.*, vol. 2, pp. 37–63, 2011.
- [25] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, 2009.
- [26] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [27] D. Rim et al., “A Stability Analysis of Neural Networks and Its Application to Tsunami Early Warning,” *J. Geophys. Res. Mach. Learn. Comput.*, 2024, doi: 10.1029/2024JH000223.
- [28] E. Juanara and C. Y. Lam, “Machine Learning Approaches for Early Warning of Tsunami Induced by Volcano Flank Collapse and Implication for Future Risk Management: Case of Anak Krakatau,” *Ocean Modell.*, vol. 194, Art. no. 102497, 2025.
- [29] S. P. D. Sriyanto, B. Adriano, Y. Fujii, and S. Koshimura, “Estimation of high-resolution tsunami inundation depth using deep learning models: Case study of Pangandaran, Indonesia,” *Ocean Eng.*, vol. 330, Art. no. 121019, 2025.
- [30] I. E. Mulia, “Machine learning approaches for tsunami early warning,” in *Probabilistic Tsunami Hazard and Risk Analysis*, Elsevier, 2025, pp. 325–337.