applied sciences

MDPI

Article

Event-Wise Validation of Machine Learning for Magnitude-Threshold Classification Using the Turkish Strong-Motion Database (SMD-TR)

Celalettin Arslan * and Faruk Baturalp Günay

Department of Computer Engineering, Faculty of Engineering, Atatürk University, 25240 Erzurum, Türkiye; baturalp@atauni.edu.tr

* Correspondence: celalettinarslan25@gmail.com

Featured Application

The workflow supports retrospective audit and calibration of magnitude-threshold classifiers in strong-motion archives. It is not an operational earthquake early-warning implementation because the predictors are full-record peak parameters.

Abstract

Magnitude-threshold classifiers trained on multiple station records from the same earthquake are vulnerable to event-level data leakage, and comparisons across magnitude boundaries can be misleading because the class definition and prevalence change. We evaluated ten machine-learning classifiers using the official SMD-TR Metadata.csv and IM_RotD50.csv files. Record metadata and orientation-independent horizontal RotD50 intensity measures were joined one-to-one by a waveform identifier (WFID), and earthquakes were grouped by the official earthquake identifier (EQID). The Mw-only cohort contained 35,198 records from 3377 events and 965 stations. Events, rather than records, were assigned to an 80% development subset and a locked 20% independent test subset. Hyperparameters, classifier, and decision-score cutoff were selected within five event-wise development folds at the prespecified Mw = 5.5 boundary using event-level F2. Linear Discriminant Analysis (LDA; lsqr solver with automatic shrinkage) was selected in development (mean F2 = 0.903, SD = 0.043; score cutoff = 0.703). On 676 independent test events, including 23 positives, LDA achieved F2 = 0.779 (95% event-bootstrap interval 0.625–0.896), recall = 0.826, precision = 0.633, balanced accuracy = 0.905, MCC = 0.712, and PR-AUC = 0.865 (TN = 642, FP = 11, FN = 4, TP = 19). An Extra Trees regression sensitivity analysis yielded event-level MAE = 0.171, RMSE = 0.229, and R^2 = 0.860. Boundaries from Mw 5.0 to 6.0 are reported as separate descriptive tasks, not as evidence for an optimal physical threshold. Because the predictors are full-record RotD50 peak parameters, the findings support offline calibration and audit rather than operational real-time early warning.

Keywords: strong-motion data; magnitude classification; event-wise validation; threshold calibration; class imbalance; SMD-TR; machine learning; regression sensitivity

Check for updates

Academic Editor: Nicholas Vassiliou Sarlis

Received: 20 July 2026
Revised: 17 August 2026
Accepted: 21 August 2026
Published: 24 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Earthquake early warning (EEW) seeks to characterize an event before damaging shaking reaches exposed locations. Operational systems therefore face two related but distinct problems: estimating source or shaking parameters from short, causally available

signal windows, and converting uncertain estimates into user-specific alert decisions [1–8]. Recent waveform-based machine-learning systems address the first problem with early P-wave or raw-waveform inputs [1–5,9–16]. Decision thresholds address the second problem and must reflect latency, false-alarm tolerance, missed-event consequences, site conditions, and the needs of the protected system [6,7,17–19].

Magnitude M = 5.5 is frequently used as a practical reference in rapid magnitude studies and application-specific warning frameworks [4,7,15,16]. It is not, however, a universally optimal alert boundary. SASMEX, intensity-based alerting studies, and risk-informed infrastructure applications illustrate that alert rules depend on distance, predicted shaking, fragility, and risk tolerance rather than magnitude alone [17–19]. A magnitude boundary is therefore a policy definition; changing it also changes the target labels and class prevalence. Predictive scores from different label definitions cannot by themselves establish that one physical boundary is superior.

A second methodological issue arises from the hierarchical structure of strong-motion databases. One earthquake may generate many station records. A random record-level split can place recordings from the same event in both training and test subsets, allowing event-specific information to cross the partition. Grouped or blocked validation is needed whenever observations share a higher-level source [20]. Model or cutoff selection on the final test subset introduces an additional optimistic bias [21]. These concerns are especially important at high magnitude boundaries, where only a few earthquakes may supply many nominally separate records.

The SMD-TR database is a national archive containing strong-motion recordings, metadata, and processed intensity measures from Türkiye [22]. The official Metadata.csv and IM_RotD50.csv files provide earthquake and station metadata together with orientation-independent horizontal RotD50 PGA, PGV, and PGD. Their explicit component convention removes ambiguity about which ground-motion component is represented. These variables support an offline classification benchmark, but full-record peaks are not equivalent to causally available features from the first seconds of an EEW sequence.

This study asks a narrower, reproducible question: how well do ten conventional classifiers distinguish events above a prespecified Mw = 5.5 boundary when validation is performed by an official earthquake identifier, model development is isolated from the final test set, and missed events receive greater weight? The contributions are: (i) an audit and one-to-one WFID joining of the official metadata and RotD50 files; (ii) an Mw-only cohort with event-wise development, cross-validation, and independent testing; (iii) complete preprocessing and hyperparameter disclosure; (iv) event-level F2, balanced accuracy, MCC, PR-AUC, baselines, cost-ratio sensitivity, and bootstrap uncertainty; and (v) continuous-magnitude regression as a sensitivity analysis. Results for Mw 5.0–6.0 remain separate descriptive tasks.

2. Materials and Methods

2.1. Study Scope and Target Definition

The analysis is retrospective and uses completed strong-motion records. For event i with reported moment magnitude Mi, the binary target at policy boundary T is $y_i(T) = 1[M_i \geq T]$. The primary boundary T = 5.5 was prespecified from its recurrent use as a reference in the literature boundaries were restricted to one-decimal increments because the catalogue magnitudes are themselves recorded to one decimal place; a cutoff reported to finer precision would have no support in the target resolution [4,7,15,16]; it was not selected by maximizing test performance. Eleven one-decimal boundaries from 5.0 to 6.0 were evaluated descriptively.

The Each boundary defines a different outcome and prevalence. Cross-boundary curves therefore describe sensitivity to the policy definition but are not used to claim an optimal physical magnitude threshold. Classifier, hyperparameter, and score-cutoff choices were made only within the development data. The independent test subset was opened after the model at M = 5.5 had been locked.

2.2. Data Source, Audit, and Primary Cohort

We obtained the official SMD-TR Metadata.csv and IM_RotD50.csv files from the DesignSafe-CI record [22]. Metadata.csv contains 95,890 rows. After removing one blank RotD50 row, IM_RotD50.csv contains 45,175 valid records. Metadata.csv includes 45,176 records for which both horizontal components are marked as processed; the sole processed horizontal metadata record without a RotD50 row is WFID 4125. It belongs to an Mw 5.8 event that remains represented by other stations and was excluded transparently. The WFID join was verified as one-to-one.

Because the source magnitudes were not homogenized by conversion equations [22], the primary cohort retained only records reported as Mw. We also required valid EQID and WFID values and positive finite Mw, epicentral distance, RotD50 PGA, RotD50 PGV, and RotD50 PGD. These rules produced 35,198 records from 3377 events and 965 stations (Table 1). No statistical imputation or positive-value outlier deletion was applied. PGA was converted from g to cm/s^2 using 980.665 cm/s^2 per g; PGV, PGD, and Repi retained their official units of cm/s, cm, and km (Table 2).

Table 1. Official SMD-TR RotD50 cohort construction and event-wise partition.

Stage	Records	Events	Stations	Purpose
Official SMD-TR metadata	95,890	9244	1022	All Metadata.csv rows
Valid RotD50 rows joined by WFID	45,175	5911	969	Complete official RotD50 PGA/PGV/PGD
Both horizontal components processed	45,175	5911	969	P/P requirement; WFID 4125 documented as absent
Final Mw RotD50 cohort	35,198	3377	965	All primary analyses
Development events	28,603	2701	953	Tuning and five-fold event CV
Independent test events	6595	676	843	Locked final evaluation

Table 2. RotD50 primary-cohort descriptive statistics at the record level.

Variable	Mean	SD	Minimum	Q1	Median	Q3	Maximum
Magnitude (Mw)	4.454	0.847	3.000	3.900	4.200	4.800	7.800
PGA-RotD50 (cm/s^2)	3.688	24.930	0.012	0.285	0.670	1.779	1356.530
PGV-RotD50 (cm/s)	0.394	3.537	<0.001	0.013	0.033	0.099	176.060
PGD-RotD50 (cm)	0.314	3.172	<0.001	0.001	0.004	0.015	138.313
Repi (km)	104.832	69.690	0.600	57.000	88.000	137.000	566.000

The three motion predictors are the official RotD50 measures calculated from the two horizontal components [22,23]. RotD50 is an orientation-independent median spectral/peak measure and is therefore neither an arbitrarily selected single component nor a maximum horizontal value. The analysis uses the supplied processed values and does not recompute intensity measures from raw waveforms. The event-level magnitude distribution of the development and locked independent-test subsets is shown in Figure 1.

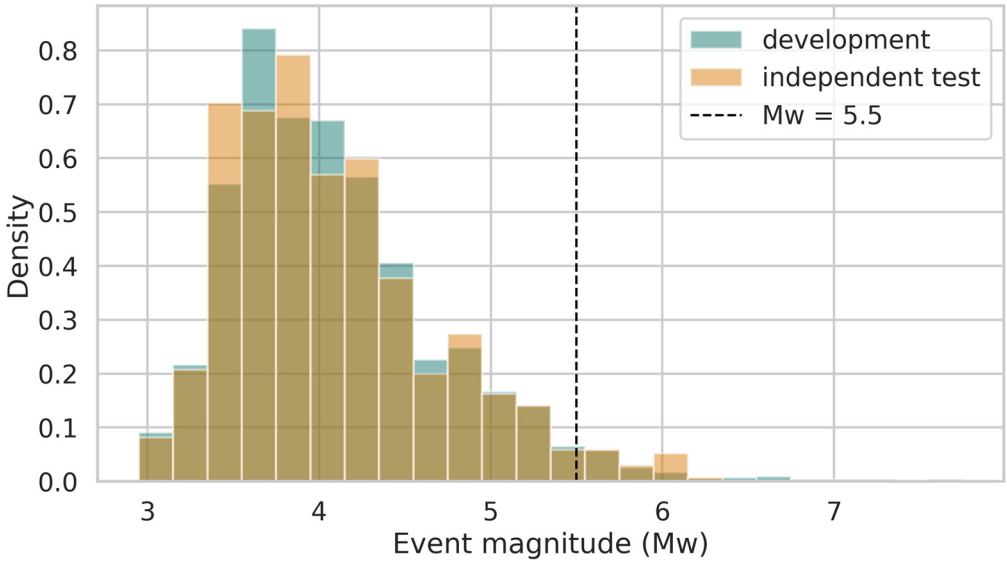


Figure 1. Event-level magnitude distribution in the official RotD50 cohort, with the development and locked independent-test subsets shown separately. The dashed line marks the prespecified Mw = 5.5 boundary.

2.3. Event Identification and Data Partitions

The official EQID was used as the grouping variable and WFID as the record-level join key. All records sharing an EQID were retained in the same subset and cross-validation fold. Magnitude was internally consistent within every retained EQID group, so no fallback event identifier was required.

EQID groups were split once into 80% development (2701 events; 28,603 records) and 20% independent test (676 events; 6595 records) using random seed 42. Stratification preserved the observed event support for <5.0, each 0.1-magnitude level from 5.0 through 5.9, and ≥6.0. There was no event overlap. The development subset was then assigned to five fixed event-wise folds. All feature, hyperparameter, classifier, and score-cutoff decisions used only these folds. At Mw = 5.5, development and test contained 87 and 23 positive events, respectively. Figure 2 summarizes the complete leakage-free workflow.

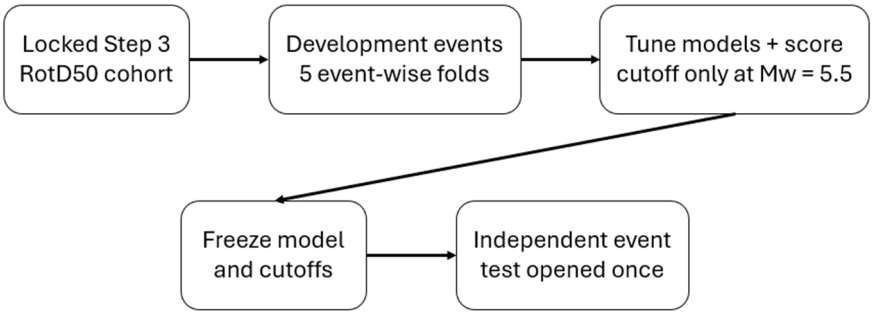


Figure 2. Leakage-free analysis workflow. Metadata and RotD50 records are joined by WFID, grouped by EQID, separated into development and locked test events, and evaluated only after development-stage selection.

2.4. Predictors and Preprocessing

The predictors were PGA-RotD50, PGV-RotD50, PGD-RotD50, and epicentral distance (Repi). Because all four variables were positive and strongly right-skewed, log10 transformation was applied within each scikit-learn (version 1.8.0, running under Python 3.12.13) pipeline. Standardization was fitted within each training fold for SVM, logistic regression, KNN, Gaussian Naive Bayes, and LDA; tree-based models used the log-transformed values

without standardization. No transformation parameter was estimated from a validation or test fold.

PGA, PGV, and PGD are full-record maxima, not short-window P-wave attributes. Repi provides source–station geometry, but the models do not separately estimate geometric-spreading or anelastic-attenuation coefficients. They are predictive classifiers, not physical ground-motion models. The effect of this design choice is considered in the interpretation and the Limitations section (Section 5).

2.5. Models, Tuning, and Event Balancing

Ten scikit-learn classifiers were compared: Random Forest (RF), Gradient Boosting (GB), Extra Trees (ET), AdaBoost (ADA), support vector classification with an RBF kernel (SVM), logistic regression (LR), k-nearest neighbors (KNN), Gaussian Naive Bayes (NB), Decision Tree (DT), and Linear Discriminant Analysis (LDA) [24]. The algorithm variants, candidate grids, and selected settings are given in Table 3; the complete machine-readable configuration and tuning log are supplied with the code.

Table 3. Candidate grids and selected classifier configurations at Mw = 5.5.

Model	Candidate Grid	Selected Setting	Scaling/Weight
RF	Depth none/12; leaf 1/4; features sqrt	300 trees; depth 12; leaf 4; sqrt	No scaling; weighted
GB	Trees 100/200; rate 0.05/0.1; depth 2	200; rate 0.1; depth 2	No scaling; weighted
ET	Depth none/12; leaf 1/4; features sqrt	300 trees; depth None; leaf 4; sqrt	No scaling; weighted
ADA	Trees 100/200; rate 0.05/0.2	200; rate 0.2	No scaling; weighted
SVM	RBF; C 0.5/2; gamma scale/0.1	RBF; C 2; gamma scale	Standardized; weighted
LR	C 0.1/1/10	C 0.1; lbfgs; max_iter 3000	Standardized; weighted
KNN	k 7/21; uniform/distance	k 21; uniform	Standardized; capped
NB	var_smoothing $10^{-11}/10^{-9}/10^{-7}$	10^{-11}	Standardized; weighted
DT	Depth 4/8/none; leaf 1/10	Depth 8; leaf 10	No scaling; weighted
LDA	svd or lsqr; shrinkage none/auto	lsqr; automatic shrinkage; equal priors	Standardized; capped

Hyperparameters were tuned at the prespecified Mw = 5.5 boundary using five event-wise development folds and mean event-level F2. To prevent earthquakes with many stations from dominating training, at most 25 records per event were retained in each training fold. Inverse-event-size and inverse-class-frequency weights were used by estimators that accept sample weights; KNN and LDA were fitted to the common capped sample because they do not accept this argument. Final classifier and score-cutoff selection were also event-level.

2.6. Event Scores, Metrics, and Uncertainty

A fitted classifier produced a score for each station record, and the event score was the arithmetic mean of all scores sharing an EQID. The decision-score cutoff was selected from out-of-fold development event scores; for the selected LDA model it was 0.703. F2 was primary because it weights recall more strongly than precision and reflects the stated preference against missed above-boundary events. We also evaluated false-negative-to-false-positive cost ratios of 2:1, 5:1, and 10:1 without using the independent test for cutoff choice.

Formally, the score of an event is the arithmetic mean of the scores of its records (Equation (1)), and, for confusion counts TP, TN, FP, and FN, the recall-weighted F-score, balanced accuracy, and Matthews correlation coefficient are defined in Equations (2)–(4):

$$s_e = \frac{1}{N_e} \sum_{i \in e} \hat{s}_i \tag{1}$$

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \,\text{Precision} + \text{Recall}}, \qquad \beta = 2 \tag{2}$$

$$\text{BA} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{3}$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

Reported event-level metrics are accuracy, balanced accuracy, precision, recall (sensitivity), specificity, F1, F2, Matthews correlation coefficient (MCC), and area under the precision–recall curve (PR-AUC). PR-AUC and MCC are included because accuracy and ROC-oriented summaries can be misleading under severe class imbalance [25,26]. Confusion counts and the majority-class accuracy baseline are reported. Uncertainty on the independent event test was estimated using 2000 nonparametric bootstrap resamples of events; 2.5th and 97.5th quantiles define 95% intervals.

2.7. Regression-First Sensitivity Analysis

To avoid training a separate target for each magnitude boundary, Random Forest, Gradient Boosting, and Extra Trees regressors were fitted to continuous Mw using the same capped, event-weighted development data. Models were compared by event-level MAE across the same five folds. Extra Trees had the lowest mean development MAE and was locked before evaluation on the independent events. Its continuous predictions were then thresholded at the prespecified Mw = 5.5 boundary. This is a sensitivity analysis rather than a physical source-magnitude model.

2.8. Reproducibility

All stochastic steps use random seed 42. The reproducibility materials include source-file checksums, the deterministic cohort-construction script, the derived-cohort checksum, the EQID split and fold manifests, executable model code, package requirements, selected hyperparameters, complete numeric outputs, event predictions, bootstrap replicates, and figure generation. The third-party Metadata.csv and IM_RotD50.csv files are obtained directly from DesignSafe-CI and are not redistributed by the authors. The Data Availability Statement provides the DOI of the code-and-results archive.

3. Results

3.1. Data Audit

The primary cohort contained 35,198 records from 3377 EQID-defined events and 965 stations. The median event contributed five records (interquartile range 2–12), and the maximum was 244, confirming that ungrouped record counts would heavily weight a small number of well-recorded earthquakes. Event-wise splitting removed direct same-earthquake partition leakage. The final cohort spans 19 August 1976 to 22 February 2023 and contains 110 events with Mw $\geq$ 5.5.

3.2. Development-Set Model Selection at M = 5.5

At the prespecified Mw = 5.5 boundary, LDA had the largest mean event-level F2 across the five development folds (0.903, SD 0.043; Table 4), narrowly followed by SVM (0.902, SD 0.044). LDA mean recall was 0.955, precision 0.747, balanced accuracy 0.972, MCC 0.838, and PR-AUC 0.931. The selected configuration used the lsqr solver with automatic shrinkage and equal class priors; its development-selected score cutoff was 0.703. The near tie with SVM limits claims of algorithmic superiority and supports interpreting selection as a deterministic development-stage decision.

Table 4. Five-fold development performance at the prespecified Mw = 5.5 boundary (event level; mean ± SD).

Model	F2	F1	Recall	Precision	Balanced Accuracy	MCC	PR-AUC
LDA	0.903 ± 0.043	0.836 ± 0.057	0.955 ± 0.047	0.747 ± 0.081	0.972 ± 0.023	0.838 ± 0.055	0.931 ± 0.030
SVM	0.902 ± 0.044	0.847 ± 0.056	0.943 ± 0.039	0.770 ± 0.074	0.967 ± 0.021	0.846 ± 0.056	0.869 ± 0.086
RF	0.892 ± 0.051	0.840 ± 0.058	0.931 ± 0.048	0.766 ± 0.067	0.961 ± 0.025	0.839 ± 0.058	0.910 ± 0.058
LR	0.889 ± 0.055	0.809 ± 0.071	0.954 ± 0.049	0.704 ± 0.092	0.970 ± 0.026	0.812 ± 0.070	0.917 ± 0.033
ET	0.888 ± 0.054	0.832 ± 0.063	0.931 ± 0.048	0.753 ± 0.076	0.960 ± 0.025	0.831 ± 0.063	0.922 ± 0.049
GB	0.882 ± 0.097	0.818 ± 0.091	0.931 ± 0.105	0.731 ± 0.088	0.960 ± 0.054	0.818 ± 0.095	0.891 ± 0.062
DT	0.878 ± 0.051	0.798 ± 0.067	0.942 ± 0.042	0.694 ± 0.083	0.964 ± 0.022	0.801 ± 0.065	0.816 ± 0.117
KNN	0.878 ± 0.062	0.773 ± 0.069	0.965 ± 0.052	0.645 ± 0.072	0.974 ± 0.028	0.781 ± 0.068	0.933 ± 0.043
ADA	0.852 ± 0.057	0.769 ± 0.058	0.919 ± 0.066	0.664 ± 0.069	0.952 ± 0.033	0.772 ± 0.058	0.877 ± 0.073
NB	0.663 ± 0.056	0.519 ± 0.037	0.816 ± 0.088	0.382 ± 0.027	0.886 ± 0.043	0.538 ± 0.045	0.537 ± 0.156

Figure 3 compares all ten classifiers at the primary Mw = 5.5 task. Figure 4 then shows LDA development F2 across the separate one-decimal boundary tasks from Mw 5.0 to 6.0. Values at different boundaries are not a ranking of magnitude policies because event support and target labels change with the boundary.

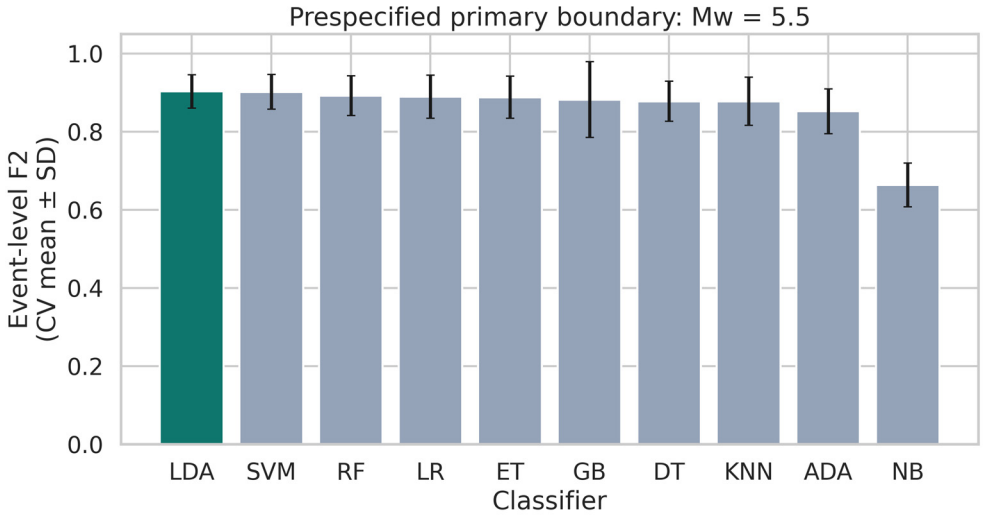


Figure 3. Five-fold development event-level F2 for all ten classifiers at the prespecified Mw = 5.5 task. LDA and SVM were nearly tied; LDA was selected by the locked development rule. The selected LDA model is shown in green, and all other classifiers are shown in gray; error bars denote the fold-wise standard deviation.

3.3. Locked Independent Event Test

The locked LDA model was evaluated once on 676 previously unseen events (Table 5). Twenty-three events met Mw ≥ 5.5. Nineteen were detected, four were missed, and 11 lower-magnitude events were false positives (TN = 642, FP = 11, FN = 4, TP = 19; Figure 5). Event-level F2 was 0.779, F1 was 0.717, recall was 0.826, precision was 0.633, balanced

accuracy was 0.905, MCC was 0.712, and PR-AUC was 0.865. Accuracy was 0.978, compared with a majority-negative accuracy baseline of 0.966.

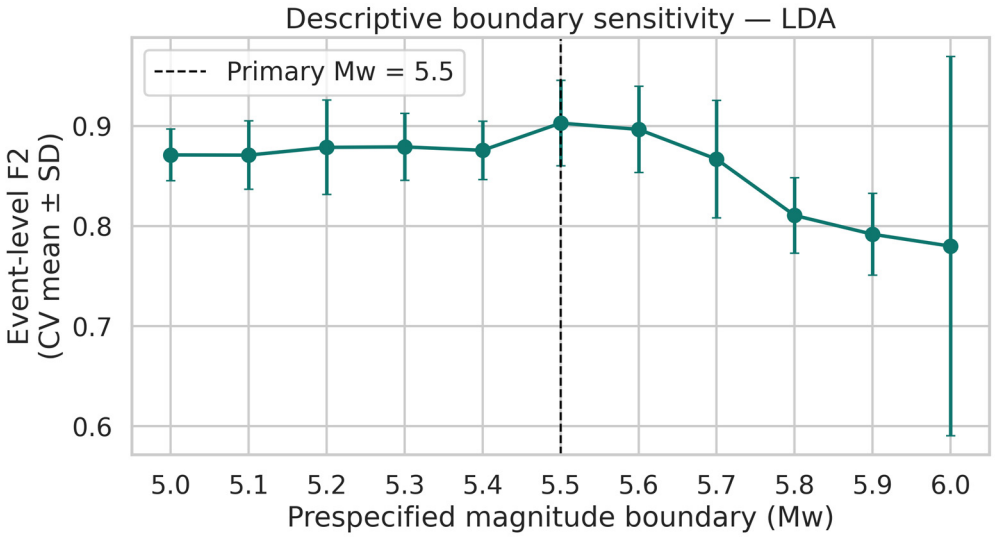


Figure 4. Five-fold development event-level F2 of the selected LDA model across separate magnitude-boundary tasks. Error bars show fold SD; the dashed line marks the prespecified Mw = 5.5 task, and boundaries are not ranked as physical policies.

Table 5. Locked independent-event performance at Mw = 5.5. LDA confidence intervals are 2000-replicate event-bootstrap intervals; regression values are descriptive point estimates.

Metric	LDA Classifier	ET Regression Then Threshold	Interpretive Note
Accuracy	0.978 [0.966, 0.988]	0.987	Majority baseline 0.966
Balanced accuracy	0.905 [0.817, 0.974]	0.846	Balances sensitivity and specificity
Precision	0.633 [0.454, 0.794]	0.889	False-alarm reliability
Recall	0.826 [0.652, 0.963]	0.696	19 of 23 positives detected
F1	0.717 [0.558, 0.836]	0.780	Equal precision/recall weight
F2	0.779 [0.625, 0.896]	0.727	LR baseline 0.806; no test reselection
MCC	0.712 [0.557, 0.833]	0.780	Imbalance-aware correlation
PR-AUC	0.865 [0.733, 0.954]	0.871	Ranking under imbalance
Confusion counts	642/11/4/19	651/2/7/16	TN/FP/FN/TP

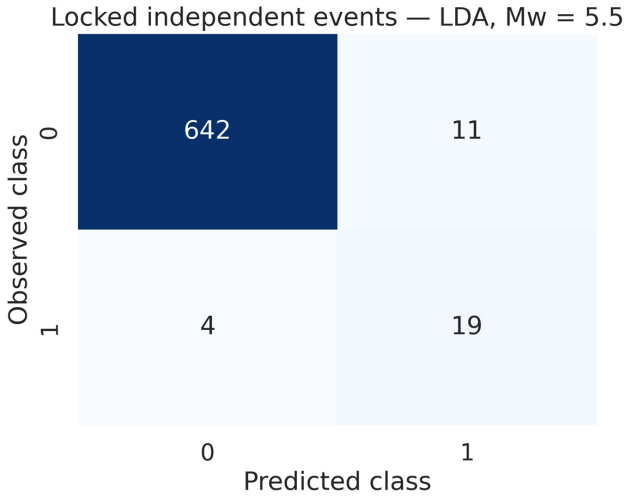


Figure 5. Confusion matrix for the development-selected LDA classifier on the locked independent event test at Mw = 5.5. Darker cell shading indicates a larger number of events in the corresponding cell.

Event-bootstrap uncertainty remained material: the 95% interval was 0.625–0.896 for F2, 0.652–0.963 for recall, 0.454–0.794 for precision, 0.817–0.974 for balanced accuracy, 0.557–0.833 for MCC, and 0.733–0.954 for PR-AUC. A logistic-regression baseline using its own development-selected cutoff obtained test F2 = 0.806. It was not substituted for LDA because changing the selected model after observing test results would invalidate the locked-test design.

3.4. Regression Sensitivity

Extra Trees had the smallest development regression error (mean event-level MAE 0.170, SD 0.006; mean RMSE 0.233; mean R^2 0.863). On the locked test events, its MAE was 0.171 magnitude units, RMSE was 0.229, and R^2 was 0.860. Applying Mw = 5.5 to the continuous predictions yielded F2 = 0.727, recall = 0.696, precision = 0.889, TN = 651, FP = 2, FN = 7, and TP = 16. This sensitivity analysis is reported without using its test result to revise classifier selection.

3.5. Predictive Feature Importance

Post-lock record-level permutation importance on records from independent test events showed the largest decrease in F2 after shuffling PGD-RotD50 (mean 0.757, SD 0.017; Figure 6), followed by Repi (0.104, SD 0.012), PGV-RotD50 (0.018, SD 0.005), and PGA-RotD50 (0.003, SD 0.004). These values describe the dependence of the fitted LDA pipeline on correlated predictors in this cohort; they were not used for feature selection and do not identify causal physical contributions.

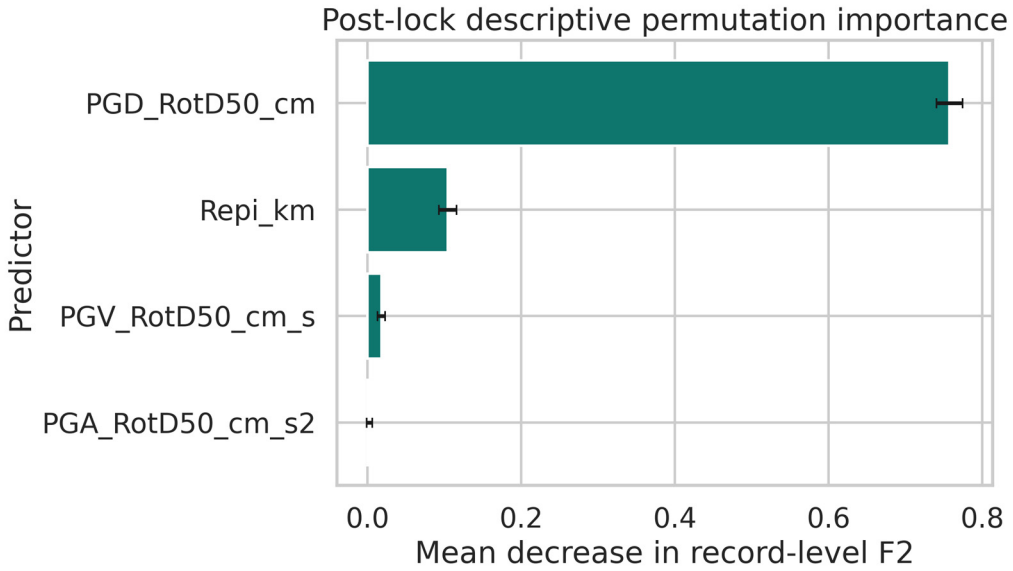


Figure 6. Post-lock record-level permutation importance of the selected LDA pipeline on records belonging to independent test events. Importance is the decrease in F2 and is descriptive rather than causal.

4. Discussion

The event-wise RotD50 analysis indicates that the data do not support an optimal Turkish magnitude boundary defined to sub-decimal precision, because catalogue magnitudes are recorded only to one decimal place. Mw = 5.5 is instead treated as an externally chosen policy reference, while the scientific result concerns event-level performance and uncertainty under a locked development-to-test protocol using official RotD50 inputs.

Event-wise partitioning is consequential because 35,198 records represent 3377 earthquakes, and a single event contributed as many as 244 records. Keeping official EQID

groups intact provides a more credible estimate of transfer to unseen earthquakes than a random station-record split [20]. Isolating 676 events for final testing further reduces selection bias [21]. Although the test now includes 23 positive events at Mw = 5.5, the bootstrap intervals still preclude fine-grained claims of superiority among near-tied models.

The use of F2 makes the risk preference explicit: a missed above-boundary event is treated as more consequential than an additional false positive. Reporting precision, recall, MCC, PR-AUC, confusion counts, and cost-ratio sensitivity exposes the trade-off rather than hiding it in accuracy or F1 [25,26]. However, the appropriate cost ratio is an application decision. A railway, hospital, industrial plant, and public alerting authority may rationally choose different rules based on downtime, fragility, and expected shaking [7,18,19]. Magnitude classification alone is therefore not an alert policy.

The regression sensitivity addresses the concern that threshold-specific labels impede comparisons across boundaries. Extra Trees achieved test MAE = 0.171 and R^2 = 0.860 in this cohort, but its thresholded F2 was 0.727 compared with 0.779 for the locked classifier. Both approaches use the same four full-record variables and national archive; regression is therefore a corroborative sensitivity analysis, not independent validation or a reason to reselect a model.

Epicentral distance and RotD50 peak motions provide a compact predictive description but only limited physical interpretability. The model does not estimate source, path, site, geometric-spreading, or anelastic-attenuation terms separately. PGD dominance in permutation importance may reflect correlation structure, scaling, the full-record processing convention, and cohort selection, as well as magnitude dependence. A physically structured extension could compare learned residuals with a ground-motion model and include site, regional, or monotonic constraints.

A defensible present application is retrospective quality assurance: a network analyst can ingest a completed event catalogue, verify units and component definitions, form event groups, run event-wise validation, inspect false negatives and false positives at a prespecified boundary, and decide whether a classifier is stable enough for further study. Translation to live EEW would require a separate end-to-end pipeline using causally truncated waveform windows, measured feature latency, evolving multi-station aggregation, site-specific shaking or impact predictions, and prospective testing. The current results do not validate those steps.

5. Limitations

Several methodological limitations arise from the chosen predictors and modeling framework. The predictors rely on full-record RotD50 peaks, meaning their availability can follow, rather than precede, damaging motion; thus, no early P-wave window, latency, packet-loss, or real-time truncation experiments were performed. Second, the results are specific to the official horizontal RotD50 intensity measures and their published processing pipeline. Other component definitions, response-spectrum periods, and raw-waveform feature constructions were not compared, and PGA was converted from g to cm/s^2 using a fixed standard-gravity factor. Additionally, the models use conventional algorithms with a deliberately restricted four-feature set that excludes raw waveforms, time–frequency attributes, site conditions, depth, focal mechanism, azimuth, network latency, and explicit physical attenuation. Consequently, the permutation importance is predictive, not causal.

Another set of limitations concerns the cohort selection and target variable uncertainties. Although 45,175 records had valid RotD50 values, restricting the analysis to reported Mw reduced the cohort to 35,198 records. This complete-measure, Mw-only selection may not represent all historical network operations or earthquakes recorded only under other magnitude scales. Furthermore, the primary analysis relies on reported Mw without quan-

tifying catalogue-magnitude uncertainty, inter-agency revisions, alternative conversion relations, or label noise near the Mw = 5.5 boundary.

Finally, the validation strategy and the nature of the operational threshold present certain constraints. The independent test is event-disjoint but not external, chronological, spatially held out, or prospectively collected. Because it contains 23 events at or above Mw = 5.5, station, network, and regional dependencies may remain, and the event bootstrap cannot create information about unobserved regions or larger earthquakes. Ultimately, M = 5.5 is a prespecified reference, not a recommended national alert rule. Real operational decisions should combine short-window source estimates with expected site shaking, uncertainty, infrastructure fragility, and stakeholder-specific false-negative and false-positive costs.

6. Conclusions

This event-wise study provides a reproducible benchmark for offline magnitude-threshold classification using official SMD-TR RotD50 intensity measures. After forming a 35,198-record Mw cohort from 3377 EQID-defined events and isolating 676 events for final testing, LDA achieved event-level F2 = 0.779, recall = 0.826, precision = 0.633, balanced accuracy = 0.905, MCC = 0.712, and PR-AUC = 0.865 at the prespecified Mw = 5.5 boundary. Extra Trees regression produced test MAE = 0.171 and R^2 = 0.860 as a sensitivity analysis.

The main methodological conclusion is that official event grouping, development-only selection, class-imbalance-aware metrics, explicit risk preferences, baselines, and event-level uncertainty materially strengthen the evidential basis of a magnitude-threshold classifier study. Because the feature set consists of full-record RotD50 peaks, the work supports retrospective calibration and audit only. Real-time EEW relevance requires prospective validation with truncated waveforms, measured latency, physical or impact-informed modeling, and external events.

Author Contributions: Conceptualization, C.A. and F.B.G.; methodology, C.A.; software, C.A.; validation, F.B.G.; formal analysis, C.A.; investigation, C.A.; resources, F.B.G.; data curation, C.A.; writing—original draft preparation, C.A.; writing—review and editing, C.A. and F.B.G.; visualization, C.A.; supervision, F.B.G.; project administration, C.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source database is SMD-TR [22], available from DesignSafe-CI at https://doi.org/10.17603/ds2-f21x-s189. The official Metadata.csv and IM_RotD50.csv files are third-party data and are not redistributed by the authors. The reproducibility materials, including the cohort-construction code, exact source and output checksums, EQID split and fold manifests, executable analysis, complete numeric outputs, and bootstrap results, are deposited at https://doi.org/10.5281/zenodo.21974715.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mousavi, S.M.; Beroza, G.C. A Machine-Learning Approach for Earthquake Magnitude Estimation. *Geophys. Res. Lett.* **2020**, *47*, e2019GL085976. [CrossRef]
2. Münchmeyer, J.; Bindi, D.; Leser, U.; Tilmann, F. Earthquake Magnitude and Location Estimation from Real-Time Seismic Waveforms with a Transformer Network. *Geophys. J. Int.* **2021**, *226*, 1086–1104. [CrossRef]
3. Lara, P.; Bletery, Q.; Ampuero, J.P.; Inza, A.; Tavera, H. Earthquake Early Warning Starting from 3 s of Records on a Single Station with Machine Learning. *J. Geophys. Res. Solid Earth* **2023**, *128*, e2023JB026575. [CrossRef]

4. Zhu, J.; Zhou, Y.; Liu, H.; Jiao, C.; Li, S.; Fan, T.; Wei, Y.; Song, J. Rapid Earthquake Magnitude Classification Using Single-Station Data Based on Machine Learning. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 7500705. [CrossRef]
5. Navarro-Rodríguez, A.; Castro-Artola, O.A.; García-Guerrero, E.E.; Aguirre-Castro, O.A.; Tamayo-Pérez, U.J.; López-Mercado, C.A.; Inzunza-Gonzalez, E. Recent Advances in Early Earthquake Magnitude Estimation by Using Machine Learning Algorithms: A Systematic Review. *Appl. Sci.* **2025**, *15*, 3492. [CrossRef]
6. Zollo, A.; Amoroso, O.; Lancieri, M.; Lomax, A.; Nielsen, S. A Threshold-Based Earthquake Early Warning Using Dense Accelerometer Networks. *Geophys. J. Int.* **2010**, *183*, 963–974. [CrossRef]
7. Zhu, J.; Sun, W.; Li, S.; Yao, K.; Song, J. Threshold-Based Earthquake Early Warning for High-Speed Railways Using Deep Learning. *Reliab. Eng. Syst. Saf.* **2024**, *250*, 110268. [CrossRef]
8. Allen, R.M.; Melgar, D. Earthquake Early Warning: Advances, Scientific Challenges, and Societal Needs. *Annu. Rev. Earth Planet. Sci.* **2019**, *47*, 361–388. [CrossRef]
9. Dybing, S.N.; Yeck, W.L.; Cole, H.M.; Melgar, D. Rapid Estimation of Single-Station Earthquake Magnitudes with Machine Learning on a Global Scale. *Bull. Seismol. Soc. Am.* **2024**, *114*, 1523–1538. [CrossRef]
10. Meng, F.; Ren, T.; Liu, Z.; Zhong, Z. Toward Earthquake Early Warning: A Convolutional Neural Network for Rapid Earthquake Magnitude Estimation. *Artif. Intell. Geosci.* **2023**, *4*, 39–44. [CrossRef]
11. Hou, B.; Zhou, Y.; Li, S.; Wei, Y.; Song, J. Real-Time Earthquake Magnitude Estimation via a Deep Learning Network Based on Waveform and Text Mixed Modal. *Earth Planets Space* **2024**, *76*, 58. [CrossRef]
12. Zhu, J.; Li, S.; Ma, Q.; He, B.; Song, J. Support Vector Machine-Based Rapid Magnitude Estimation Using Transfer Learning for the Sichuan–Yunnan Region, China. *Bull. Seismol. Soc. Am.* **2022**, *112*, 894–904. [CrossRef]
13. Ren, T.; Liu, X.; Chen, H.; Dimirovski, G.M.; Meng, F.; Wang, P.; Zhong, Z.; Ma, Y. Seismic Severity Estimation Using Convolutional Neural Network for Earthquake Early Warning. *Geophys. J. Int.* **2023**, *234*, 1355–1362. [CrossRef]
14. Chakraborty, M.; Li, W.; Faber, J.; Rümpker, G.; Stoecker, H.; Srivastava, N. A Study on the Effect of Input Data Length on a Deep-Learning-Based Magnitude Classifier. *Solid Earth* **2022**, *13*, 1721–1729. [CrossRef]
15. Zhu, J.; Li, S.; Song, J. Multimodal Deep Learning Network for Fast Seismic Discrimination and Magnitude Classification. *Geosci. Lett.* **2025**, *12*, 39. [CrossRef]
16. Zheng, Z.; Chen, X.; Lin, B.; Jin, Z.; Jin, X.; Yu, J. A Multimodal Deep Learning Framework for Rapid Real-Time Earthquake Magnitude Classification: Applications to Earthquake Early Warning. *Bull. Seismol. Soc. Am.* **2025**, *115*, 2202–2218. [CrossRef]
17. Suárez, G. The Seismic Early Warning System of Mexico (SASMEX): A Retrospective View and Future Challenges. *Front. Earth Sci.* **2022**, *10*, 827236. [CrossRef]
18. Saunders, J.K.; Minson, S.E.; Baltay, A.S. How Low Should We Alert? Quantifying Intensity Threshold Alerting Strategies for Earthquake Early Warning in the United States. *Earth's Future* **2022**, *10*, e2021EF002515. [CrossRef]
19. Cremen, G.; Bozzoni, F.; Pistorio, S.; Galasso, C. Developing a Risk-Informed Decision-Support System for Earthquake Early Warning at a Critical Seaport. *Reliab. Eng. Syst. Saf.* **2022**, *218*, 108035. [CrossRef]
20. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
21. Varma, S.; Simon, R. Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinform.* **2006**, *7*, 91. [CrossRef] [PubMed]
22. Sandıkkaya, M.A.; Güryuva, B.; Kale, Ö.; Okçu, O.; İçen, A.; Yenier, E.; Akkar, S. An Updated Strong-Motion Database of Türkiye (SMD-TR). *Earthq. Spectra* **2024**, *40*, 847–870. [CrossRef]
23. Boore, D.M. Orientation-Independent, Nongeometric-Mean Measures of Seismic Intensity from Two Horizontal Components of Motion. *Bull. Seismol. Soc. Am.* **2010**, *100*, 1830–1835. [CrossRef]
24. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
25. Saito, T.; Rehmsmeier, M. The Precision–Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
26. Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.